Trained quantum neural networks are Gaussian processes

Filippo Girardi[1,2,3*] and Giacomo De Palma[4]

[1]Korteweg–de Vries Institute for Mathematics, University of Amsterdam, Science Park 105-107, Amsterdam, 1098 XG, The Netherlands.
[2]QuSoft, Science Park 123, Amsterdam, 1098 XG, The Netherlands.
[3]Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126, Pisa (PI), Italy.
[4]Department of Mathematics, University of Bologna, Piazza di Porta San Donato 5, 40126, Bologna (BO), Italy.

*Corresponding author(s). E-mail(s): filippo.girardi@sns.it;
Contributing authors: giacomo.depalma@unibo.it;

Abstract

We study quantum neural networks made by parametric one-qubit gates and fixed two-qubit gates in the limit of infinite width, where the generated function is the expectation value of the sum of single-qubit observables over all the qubits. First, we prove that the probability distribution of the function generated by the untrained network with randomly initialized parameters converges in distribution to a Gaussian process whenever each measured qubit is correlated only with few other measured qubits. Then, we analytically characterize the training of the network via gradient descent with square loss on supervised learning problems. We prove that, as long as the network is not affected by barren plateaus, the trained network can perfectly fit the training set and that the probability distribution of the function generated after training still converges in distribution to a Gaussian process. Finally, we consider the statistical noise of the measurement at the output of the network and prove that a polynomial number of measurements is sufficient for all the previous results to hold and that the network can always be trained in polynomial time.

Keywords: quantum machine learning, quantum neural networks, Gaussian processes

Contents

1 Introduction **3**
1.1 Our results . . . 4

2 Quantum neural networks and light cones **6**
2.1 The dataset of examples . . . 7
2.2 The model function: a quantum circuit . . . 7
2.2.1 Layers, parameters and observables . . . 7
2.2.2 The encoding of the input . . . 10
2.3 Light cones . . . 11
2.3.1 Light cones and their extensions . . . 11
2.3.2 The circuit seen by a local observable . . . 13
2.3.3 Light cones, local Hilbert spaces and classical simulability . . . 15
2.3.4 Architecture-independent bounds and geometrically local circuits . . . 17
2.4 Random initialization of quantum circuits . . . 17

2.5 Some examples and the estimate of $N(m)$ . . . 19

3 Quantum neural networks with random parameters are Gaussian processes 21
3.1 Central limit theorem and generalizations . . . 22
3.1.1 Gaussian processes and the central limit theorem . . . 22
3.1.2 Cumulants and dependency graphs . . . 24
3.2 Convergence to a Gaussian process . . . 26
3.3 Proof of Theorem 3.7 . . . 28

4 Trained quantum neural networks are Gaussian processes 31
4.1 Introduction . . . 31
4.1.1 Training quantum neural networks . . . 31
4.1.2 Trainability: the loss landscape and barren plateaus . . . 32
4.1.3 Lazy training in the quantum setting . . . 35
4.2 Gradient flow and neural tangent kernel . . . 37
4.3 Assumptions and properties of the NTK . . . 38
4.4 Analytic solution for the linearized model . . . 40
4.5 Convergence to the linearized model . . . 41
4.6 Some useful lemmas . . . 44
4.6.1 Loss function at initialization . . . 44
4.6.2 Lipschitzness of the model . . . 46
4.6.3 Original model, linearized version and lazy training . . . 50
4.6.4 The minimum of the loss for the linear model and the convergence of the kernel . . . 51
4.7 Proof of Lemma 4.5 . . . 54
4.8 Proof of Theorem 4.7 . . . 56
4.9 Proof of Corollaries 4.8 and 4.9 . . . 58
4.9.1 Proof of Corollary 4.8 . . . 59
4.10 Proof of Theorem 4.10 . . . 60
4.11 Proof of Corollary 4.11 and Theorem 4.12 . . . 61
4.11.1 Proof of Corollary 4.11 . . . 61
4.12 Proof of Lemma 4.13 . . . 62
4.13 Proof of Theorem 4.14 . . . 65
4.14 Proof of Theorem 4.15 . . . 65

5 Noisy training of quantum neural networks 67
5.1 Gradient descent with a finite number of measurements . . . 67
5.2 Deterministic gradient descent . . . 68
5.3 Stochastic gradient descent . . . 69
5.4 Proof of Theorem 5.1 . . . 74
5.5 Proof of Theorem 5.4 . . . 79
5.6 Proof of Lemma 5.6 . . . 85
5.7 Proof of Theorem 5.8 . . . 93
5.8 Proof of Theorem 5.9 . . . 94

6 The case of an infinite input space 99

7 Conclusions 106

A A counterexample for large light cones 107

B Architecture-independent bounds for light cones 109

C Computation of the analytic NTK 110

1 Introduction

The last years have witnessed the surge of artificial intelligence powered by deep neural networks, which have achieved an enormous success in several fields such as speech recognition, computer vision and natural language processing, up to the recent development of large language models such as GPT-4 [1–6]. Quantum neural networks constitute a quantum counterpart of deep neural networks. The functions that they generate are the expectation value of a quantum observable (such as the number of qubits in the 1 state) measured on the output of a quantum circuit made by parametric one-qubit and two-qubit gates. The parameters are usually trained variationally with stochastic gradient descent. Quantum neural networks have the promise to add the power of quantum computers to the capabilities of deep neural networks [7, 8]. Indeed, embedding classical data into a Hilbert space whose dimension is exponential in the number of qubits may give access to computations that are classically hard to perform [9, 10]. In particular, [11] provided an example of a family of datasets – based on the discrete logarithm – that only a quantum classifier can learn efficiently. However, in all the known examples where quantum neural networks show quantum advantage, the right parameters need to be known a priori and put in by hand, and it is unknown whether they could be found variationally. Indeed, there is no problem of practical interest yet where variationally trained quantum neural networks can provably outperform classical neural networks. Furthermore, the training of quantum neural networks can suffer from bad local minima [12, 13] or gradients whose size decreases exponentially with the number of qubits [14], a phenomenon called "barren plateaus". The possible solutions that have been proposed to avoid the issue of barren plateaus consist in changing the cost function of the network. Ref. [15] considered untrained quantum neural networks with random parameters and proved that local cost functions (*i.e.*, cost functions that are made by the sum of local terms acting on few qubits each) do not suffer from the issue, provided that the depth of the network is at most logarithmic in the number of qubits [16]. Ref. [17] proposed more general cost functions that arise from the quantum generalization of optimal mass transport and of the Lipschitz constant of [18] and are robust with respect to quantum operations acting on few qubits. However, a rigorous proof of the efficacy of these solutions is still lacking, and it has been conjectured that all the architectures for quantum neural networks that do not suffer from barren plateaus can be simulated efficiently on a classical computer [19].

A fundamental breakthrough in the mathematical understanding of the functioning of classical deep neural networks has been the proof that, in the limit of infinite width, the training is always able to perfectly fit the training examples while at the same time avoiding overfitting. More precisely, a series of works [20–26] has proved that in the limit of infinite width, the probability distribution of the function generated by a deep neural network trained with stochastic gradient descent on a supervised learning problem converges to a Gaussian process. Such Gaussian process always perfectly fits the training examples, and its mean and covariance can be computed analytically given the training data and the architecture of the network [27–35]. These results prove that the limit of infinitely many parameters is smooth and analytically solvable, and provide the first rigorous mathematical foundation to classical machine learning. The proof of these results is based on the following intermediate steps:

- The probability distribution of the function generated by an untrained deep neural network with random parameters converges to a Gaussian process;
- The variation of any single parameter during the whole training is infinitesimal, but since the number of parameters is infinite, all the infinitesimal variations add up to a finite change in the generated function. Such regime is called "lazy training".

This classical breakthrough has stimulated the study of overparameterized quantum neural networks, which has been initiated in [36], while their lazy-training regime has been studied in [37]. Ref. [38] proved that the training of quantum neural networks with constant depth happens in the lazy regime in the limit of infinite width and is capable of perfectly fitting the training examples. However, quantum neural networks with constant depth do not have any hope of quantum advantage since the past light-cone of each measured qubit has constant size and its classical simulation is efficient. Ref. [39] considers the opposite limit, where the depth is high enough such that the number of parameters grows exponentially with the number of qubits and the network can reproduce any unitary, and proves the lazy training and the trainability of the network in such a regime. Ref. [40] proves that, in the same regime, the probability distribution of the function generated at initialization converges to a Gaussian process.

1.1 Our results

In this paper, we generalize to the quantum setting the breakthrough in classical machine learning presented above. We consider quantum neural networks trained on supervised learning problems where the generated function is the expectation value of the sum of single-qubit observables over all the qubits[1] and prove for the first time trainability in the limit of infinite width in any regime where the depth is allowed to grow with the number of qubits as long as barren plateaus do not arise. To be precise, our conclusions are still valid even if only a fraction of the derivatives of the model function is exponentially suppressed, while the norm of the gradient (or, equivalently, the covariance of the model function at initialization, as in Lemma 4.30) is not exponentially suppressed. We stress that quantum circuits that are deep enough to forbid a naive efficient classical simulation but not too deep such that they do not suffer from barren plateaus do exist. Indeed, barren plateaus are not present whenever the measured observable is local and the depth of the circuit scales logarithmically with the number of qubits [16]. Such depth still allows the size of the past light-cone of each measured qubit to scale as a power of the number of qubits (see subsection 2.5 for more details). In this case, a naïve classical simulation would require exponential time.

Our first result states that the probability distribution of the function generated by a randomly initialized quantum neural network in the limit of infinite width converges in distribution to a Gaussian process when the parameters on which each measured qubit depends influence only a small number of other measured qubits (Theorem 3.7).

We then consider quantum neural networks trained in continuous time with gradient flow, and prove that the training happens in the lazy regime and is able to perfectly fit the training set (Theorem 4.10). Contrarily to [38], our results allow for a depth growing with the number of qubits and are thus valid in regimes with hope of quantum advantage (see subsubsection 4.1.3 for a detailed comparison). Furthermore, we prove that the dependence of the model on the parameters can be approximated with its linearized version around the value of the parameters at initialization (Theorems 4.12 and 4.14). The corresponding linearized evolution equation has an analytical solution that determines the probability distribution of the function generated by the trained network, which is a Gaussian process whose mean and covariance can be computed analytically (Corollary 4.9). Building on such results, we prove that the probability distribution of the function generated by the trained network converges in distribution to the Gaussian process above (Theorem 4.15).

We then focus on the issues related to real physical setting of the training of the network. On the one hand, we need to consider that the training takes place in discrete time steps, and we discuss the convergence of the gradient descent. On the other hand, we have to take into account that the function generated by a quantum neural network is the expectation value of a quantum observable, which can only be estimated via measurements. In this regard, we prove that a number of measurement polynomial in the number of qubits is enough to ensure all the previous results (Theorems 5.4 and 5.9). In particular, the probability distribution of the function generated by the quantum neural network converges in distribution to a Gaussian process also in the presence of statistical noise. The above results on the evolution of quantum neural networks in the noisy setting are our most important achievement. Indeed, they prove the trainability of the model in polynomial time with respect to the number of qubits, which is a necessary condition for quantum advantage, and prove that the limit of infinitely many parameters is smooth. Furthermore, thanks to the above results, the Seeger's generalization bounds for Gaussian processes [42] can be applied to any trained quantum neural network satisfying the hypotheses of our paper in the limit of infinite width. In particular, our results prove that such limit does not trivialize the generalization properties of quantum neural networks. Indeed, Seeger's generalization bound can be nontrivial even if, being infinitely wide, the network has infinitely many parameters with a finite training set[2].

Finally, differently from the main works which have studied the initialization of quantum neural networks [40, 43], we go beyond the standard assumption that the feature space is a finite set. Indeed, we rigorously prove that the convergence to a Gaussian process, both at initialization and during the

[1] Our model differs from [41], where a single single-qubit observable, or more generally an observable with eigenvalues ± 1, is chosen: in that setting the convergence of the training has been proved to be slower than in the classical case. In our case, on the contrary, the convergence is exponential in the training time.

[2] Of course, depending on the specific training set and on the specific neural tangent kernel of the considered quantum architecture, Seeger's generalization bound can become trivial, but this does not happen systematically for the only fact that the parameters are more than the training examples.

Table 1 Notation concerning the properties of the circuit at initialization.

Symbol	Description	Introduced in
m	number of qubits in the parameterized quantum circuit	subsubsection 2.2.1
L	number of layers in the parameterized quantum circuit	Def. 2.2
$[\ell\, m]$	layer-qubit representation	Def. 2.3
$\Theta \in \mathscr{P}$	vector of the parameters of the quantum circuit, belonging to the parameter space; typically $\mathscr{P} = [0, \pi]^{Lm}$	subsubsection 2.2.1
$\lvert\Theta\rvert$	number of parameters $\lvert\Theta\rvert := \dim \mathscr{P} = Lm$	*ibid*
$x \in \mathcal{X}$	input vector, belonging to the input (or feature) space $\mathcal{X}$; typically $\mathcal{X} = [0, \pi]^{\dim \mathcal{X}}$	*ibid.*
$U(\Theta, x)$	parameterized quantum circuit (unitary operator)	*ibid.*
$f(\Theta, x)$	function generated by the quantum neural network, often called *original model*	*ibid.*
$N(m)$	normalization factor of the model	*ibid.*
$\mathcal{M}_i$	(extended) future light cone of the parameter i	Def. 2.4, Def. 2.5 and Cor. 2.5
$\mathcal{N}_k$	(extended) past light cone of the observable k	
$\lvert\mathcal{M}\rvert$	maximal cardinality of a future light cone in the circuit	Def. 2.6
$\lvert\mathcal{N}\rvert$	maximal cardinality of a past light cone in the circuit	*ibid.*

training, is valid also when we consider an infinite feature space (Theorem 6.4). This achievement not only makes the mathematical treatment of our results complete, but also gives a precise physical insight: our proof ensures the existence of a lower bound to the rate of convergence to the Gaussian process which does not depend on the particular discretization of the feature space. In particular, the rate of convergence will not trivialize to zero as the the discretization is refined.

Summarizing, this paper identifies the limit of infinite width as a regime in which quantum neural networks are provably efficiently trainable whenever they do not suffer from barren plateaus. Therefore, despite results such as [12, 13] prove that problems other than barren plateaus can affect the trainability of variational quantum algorithms, we prove that barren plateaus are the only obstacle that can affect the trainability of quantum neural networks (at least in the limit of infinite width). Therefore, our results constitute the first step to prove that useful and provably trainable quantum neural networks do exist.

Having proved that sufficiently wide quantum neural networks that satisfy our hypotheses are efficiently trainable, we hope that our paper will stimulate the practitioners of quantum machine learning to consider wide architectures. We stress that such architectures do not actually require a quantum computer with as many qubits as their width. Indeed, as we explain in more detail in subsubsection 2.3.3 (see, in particular, Remark 2.9), the expectation value of each measured one-qubit observable can be estimated independently, and such estimate requires only the qubits in the past light-cone of the measured qubit. We further stress that our hypotheses are extremely general, and any sequence of quantum neural networks with increasing width and with not too large depth will satisfy them. What are the best wide architectures for a given supervised learning problem is a fundamental question that will need to be answered by experiments.

The article is organized as follows.

In section 2, we fix the setup and the notation of the paper. In section 3, we prove that the function generated by a quantum neural network with random parameters in the limit of infinite width is a Gaussian process. In section 4, we consider quantum neural networks trained in continuous time and prove that the trained model is able to perfectly fit the training set and that the probability distribution of the generated function generated by the trained network is a Gaussian process. In section 5, we consider quantum neural networks trained with discrete gradient descent taking into account the statistical noise, and prove that all the previous results are still valid. In section 6, we extend our result to the general case of an infinite input space. We conclude in section 7. In Appendix A we show, by means of a counterexample, that a circuit violating our hypotheses can generate a function that is not distributed as a Gaussian process. In Appendix B, we compute some architecture-independent bounds for the cardinalities of the light cones.

In order to improve the readability of the paper, section 3, section 4, section 5 and section 6 start with the presentation of the results, whose proofs are deferred to the second part of the section.

Table 2 Notation concerning the training of the circuit.

Symbol	Description	Introduced in
n	number of examples in the training set	subsection 2.1
$\mathcal{D}$	training set, whose elements are denoted by $(x^{(i)}, y^{(i)})$ for $i = 1, \dots, n$	*ibid.*
X	vector containing the inputs of the training set	*ibid.*
Y	vector containing the outputs of the training set corresponding to X	*ibid.*
$f^{\text{lin}}(\Theta, x)$	linearized model	subsection 4.4
$\hat{K}_\Theta(x, x')$	empirical neural tangent kernel	Def. 4.1
$N_K(m)$	normalization factor of the empirical neural tangent kernel	*ibid.*
$K(x, x')$	analytic neural tangent kernel	Def. 4.2
$\bar{K}(x, x')$	(uniform) limit of the analytic neural tangent kernel in the limit of infinitely many qubits	Ass. 4.3
$\bar{K}$	abbreviation for $\bar{K}(X, X^T)$	*ibid.*
λ^K_{min}	smallest eigenvalue of $\bar{K}$	*ibid.*

Table 3 Notation concerning the training of the circuit (continuation).

Symbol	Description	Introduced in
t	continuous or discrete training time	subsubsection 4.1.1
$F(t)$	vector containing the model function evaluated in the inputs of the dataset, i.e., $F(t) = f(\Theta_t, X)$	subsection 4.2
$F^{\text{lin}}(t)$	vector containing the linearized model function evaluated in the inputs of the dataset, i.e., $F^{\text{lin}}(t) = f^{\text{lin}}(\Theta^{\text{lin}}_t, X)$	subsection 4.4
η	learning rate, which enters the gradient flow equation and is a function of m	subsubsection 4.1.1 and Ass. 4.6
η_0	resized learning rate (which does not depend on m)	Ass. 4.6
$\mathcal{L}(\Theta)$	cost function for the original model according to the training set	Def. 2.1 and subsection 4.4
Θ_t	parameter vector evolving via gradient flow (or gradient descent) according to the cost function	subsubsection 4.1.1
$\mathcal{L}^{\text{lin}}(\Theta)$	cost function for the linearized model, sometimes called *linearized cost function*	subsection 4.4
Θ^{lin}_t	parameter vector evolving via gradient flow (or descent) according to the linearized cost function	*ibid.*

2 Quantum neural networks and light cones

Quantum neural networks constitute the main paradigm in quantum machine learning [7]. In order to give a quantitative description of their behaviour at initialization and during training, in this section we will fix a rather generic architecture (see Figure 2), which covers a very large class of situations of interest. Moreover, most of our results can be straightforwardly generalized to slightly different architectures[3].

In the whole paper, we consider a sequence of quantum neural networks with increasing number of qubits m, and we study the limit of infinite width $m \to \infty$. The key point of this section is to identify some global quantities which describe the physical features of this generic architecture. We will state all our results in terms of such quantities, without introducing further assumptions on the circuit. Conversely, if one wants to check if a specific circuit satisfies our theorems, only such global quantities have to be computed[4]:

$$\begin{aligned} m &= \text{ number of qubits in the circuit,} \\ L &= \text{ number of layers of the circuit,} \\ |\mathcal{M}| &= \text{ maximal dimension of the future light cone of a parameter,} \\ |\mathcal{N}| &= \text{ maximal dimension of the past light cone of an observable,} \\ N(m) &= \text{ normalization constant of the model.} \end{aligned}$$

[3] For instance, one could consider two-qubit parameterized gates instead of one-qubit ones, or consider local observable acting on $O(1)$ qubits instead of one-qubit observables. This would produce different constants in the bounds we will present.

[4] If, on the one hand, dealing only with global quantities that ignore the detailed structure is a great simplification, on the other hand we should not deduce that all the circuits which are indistinguishable in these quantities can be considered equivalent for the machine-learning problem. Indeed, the specific structure of the circuit typically may improve (or hinder) the learning process and the generalization capacity of the model.

All these quantities may depend on the numer of qubits m and, as we will discuss in subsection 2.5, their mutual growth as $m \to \infty$ will control the good behaviour of the circuit. The physical meaning of the light cones is depicted in Figure 5, Figure 6 and Figure 8.

The main notations used in this article are summarized in Table 1, Table 2 and Table 3.

2.1 The dataset of examples

Let $\mathcal{X} \subseteq \mathbb{R}^{\dim \mathcal{X}}$ be the feature space (or input space) and let $\mathcal{Y} \subseteq \mathbb{R}$ be the output space. Given a training set $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1,\dots,n} \subseteq \mathcal{X} \times \mathcal{Y}$, we will call $n = |\mathcal{D}|$ the number of examples and we will represent it in a vectorized form as follows

$$X = \begin{pmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(n)} \end{pmatrix}, \qquad Y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{pmatrix}. \tag{2.1}$$

Given any function $g : \mathbb{R} \to \mathbb{R}$, we will often use the following notation:

$$g(X) \coloneqq \begin{pmatrix} g(x^{(1)}) \\ g(x^{(2)}) \\ \vdots \\ g(x^{(n)}) \end{pmatrix}, \qquad g(X^T) \coloneqq \begin{pmatrix} g(x^{(1)}) & g(x^{(2)}) & \cdots & g(x^{(n)}) \end{pmatrix} \tag{2.2}$$

Similarly, for any bivariate function $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ we will write $K(X, X^T)$ to indicate the $n \times n$ matrix with entries $\big(K(X, X^T)\big)_{ij} \coloneqq K(x^{(i)}, x^{(j)})$ for $1 \leq i, j \leq n$.

Assumption 2.1 *We assume that the set of the possible inputs $\mathcal{X}$ is finite.*

Remark 2.1 (The case of infinite input space) In a concrete setting we will typically have a finite amount of digits to codify our inputs (over a bounded interval), so it is reasonable to suppose $\mathcal{X}$ to be finite. Even though we will present all our results under the assumption that $\mathcal{X}$ is finite – this allows to provide simpler and less technical proofs – all the final statements holds even when $\mathcal{X}$ is infinite. The reader interested in this technical treatment of the infinite case can consult section 6, where we provide complete proofs for this more general setting.

Definition 2.1 (Mean squared error) As a cost function associated to the training set $\mathcal{D}$, we will consider the *mean squared error*:

$$\begin{aligned} \mathcal{L}(\Theta) &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \Big(f(\Theta, x^{(i)}) - y^{(i)} \Big)^2 \\ &= \frac{1}{2n} \| f(X) - Y \|_2^2. \end{aligned} \tag{2.3}$$

2.2 The model function: a quantum circuit

2.2.1 Layers, parameters and observables

Following the notations of [38], let L be the number of parameterized layers of the circuit, which might depend on the number of qubits m.
Let $\theta_1, \dots, \theta_{Lm}$ be the parameters of the circuit, which we will compactly write in the vector

$$\Theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_{Lm} \end{pmatrix}, \tag{2.4}$$

whose dimension will often be denoted as $|\Theta|$. A circuit is a finite composition of parameterized layers. We need to clearly define the structure of a layer and how the parameterized gates are arranged.

Definition 2.2 A *layer* is the unitary operation $U_\ell(\Theta, x) \in \mathcal{L}(\mathcal{H})$ $(1 \leq \ell \leq L)$ resulting from

1. the application on each qubit of a different parameterized single-qubit gate $W_i(\Theta) \in \mathcal{L}(\mathbb{C}^2)$; each parameterized gate depends on a single parameter θ_i, which is different for each gate,

followed by

2. a set of one-qubit and two-qubit gates acting on disjoint qubits, *i.e.*, each qubit can be acted on by at most one gate; each gate may depend only on the input x; the resulting unitary operation will be called $V_\ell(x) \in \mathcal{L}(\mathcal{H})$.

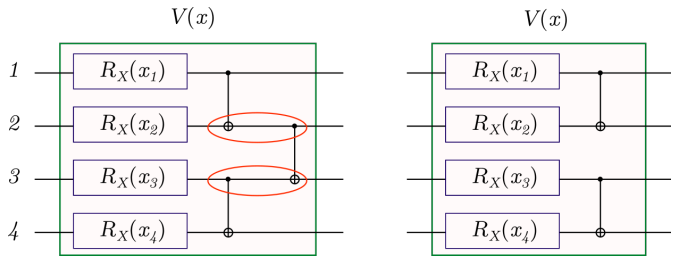


Fig. 1 On the left, an internal structure of $V(x)$ which is forbidden according to Definition 2.2: the entangling gate acting on the qubits 2 and 3 must be moved to a new layer. The structure on the right is allowed.

To clarify the second point, it is useful to look at Figure 1 as an example: the internal structure of $V(x)$ on the left is not allowed according to our definition of layer, since qubit 3 interacts with both qubit 2 and qubit 4. On the contrary, the structure on the right is allowed, since each qubit interacts with at most one different qubit.

In Figure 2 a circuit composed of L layer is represented.

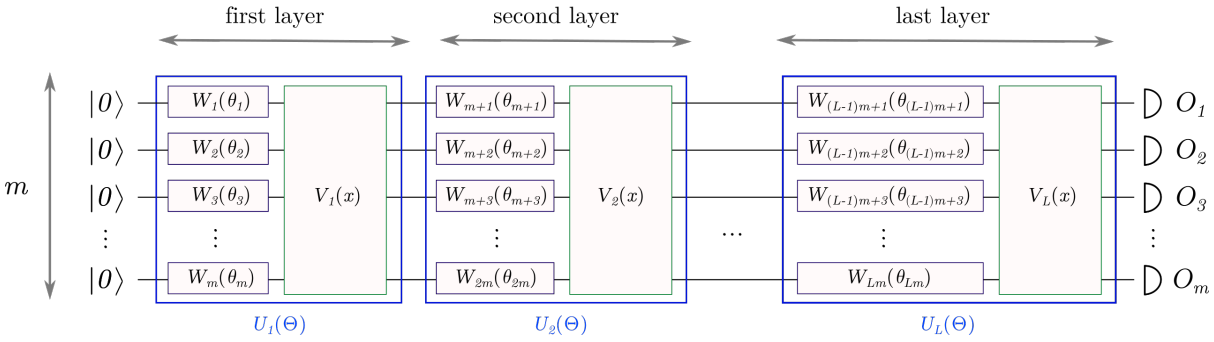


Fig. 2 Our parameterized quantum circuit.

We introduce a convenient notation for the indices of the parameters.

Definition 2.3 (Layer-qubit representation) Each parameter index $i \in \{1, \ldots, Lm\}$ can be represented in the form $i = m(\ell - 1) + k$, with $\ell \in \{1, \ldots, L\}, k \in \{1, \ldots, m\}$. k refers to the qubit involved in the single-qubit gate parameterized by θ_i, while ℓ refers to the layer in which such gate acts, as in Figure 3. The following compact notation, which we call *layer-qubit representation* of the parameter index i, simplifies the above form:

$$i = [\ell m] \equiv m(\ell - 1) + k. \tag{2.5}$$

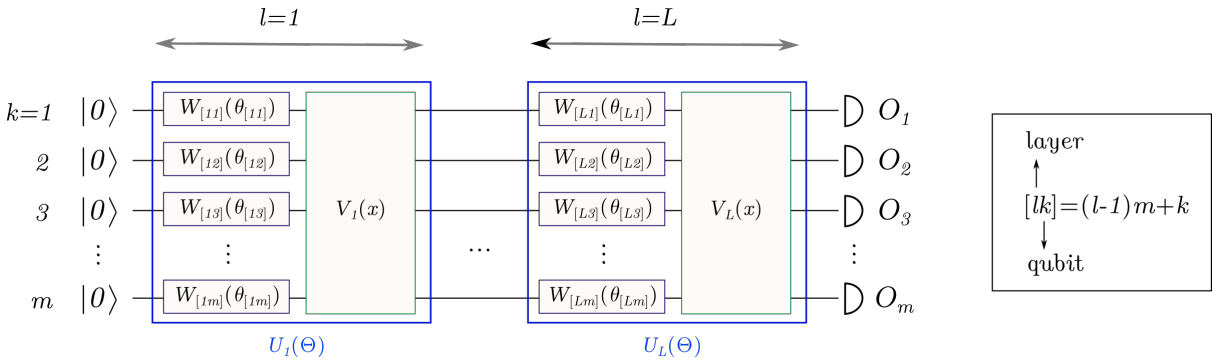


Fig. 3 The layer-qubit representation.

Therefore, a layer U_ℓ can be written as

$$\begin{aligned} U_\ell(\Theta, x) &= V_\ell(x)\left(W_{[\ell 1]} \otimes W_{[\ell 2]} \otimes \cdots \otimes W_{[\ell m]}\right)(\Theta) \\ &= V_\ell(x) W_\ell(\Theta), \end{aligned} \tag{2.6}$$

where

$$W_\ell(\Theta) = \left(W_{[\ell 1]} \otimes W_{[\ell 2]} \otimes \cdots \otimes W_{[\ell m]}\right)(\Theta). \tag{2.7}$$

The action of the circuit on the initial state $|\psi_0\rangle$ is described by the unitary operation

$$U(\Theta, x) = U_L(\Theta, x) \cdots U_2(\Theta, x) U_1(\Theta, x), \qquad |\psi_{out}\rangle = U(\Theta, x)\,|\psi_0\rangle\,. \tag{2.8}$$

Let $\mathcal{O} \in \mathcal{L}(\mathcal{H})$ be the observable of the circuit, i.e., the measurement we are interested to perform on $|\psi_{out}\rangle$.

Assumption 2.2 *$\mathcal{O}$ is a local observable given by the sum of single qubit observables O_k:*

$$\begin{aligned} \mathcal{O} = \sum_{k=1}^{m} \mathcal{O}_k = O_1 \otimes \mathbb{1}_2 \otimes \cdots \otimes \mathbb{1}_m \\ + \mathbb{1}_1 \otimes O_2 \otimes \mathbb{1}_3 \otimes \cdots \otimes \mathbb{1}_m \\ + \cdots + \\ + \mathbb{1}_1 \otimes \cdots \otimes \mathbb{1}_{m-1} \otimes O_m. \end{aligned} \tag{2.9}$$

For each k, we assume that $\mathrm{Tr}[O_k] = 0$ and that the spectrum of O_k is contained in the interval $[-1, +1]$, as in the case of Pauli observables. We further assume that the parameterized single qubit gates of the circuit $W_i(\theta_i)$ can be written in terms of hermitian operators $\mathcal{G}_i$ with spectrum $\{-1, +1\}$:

$$W_i(\theta_i) = e^{-i\mathcal{G}_i\theta_i}. \tag{2.10}$$

Therefore, the function $f(\,\cdot\,, x)$ is componentwise periodic with period π for any $x \in \mathcal{X}$.

Remark 2.2 Whenever the spectrum of $\mathcal{G}_i$ is $\{-1, +1\}$ for any $i = 1, \ldots, Lm$ (as in the case of Pauli rotation gates), in the basis diagonalizing $\mathcal{G}_i$ we have that

$$W_i(\theta_i) = \exp\left[-i \begin{pmatrix} 1 & \\ & -1 \end{pmatrix} \theta_i\right] = \begin{pmatrix} e^{-i\theta_i} & \\ & e^{i\theta_i} \end{pmatrix} \tag{2.11}$$

is a periodic function of θ_i, with period[5] π. Therefore, the domain of Θ can be identified with the compact set $\mathscr{P} = [0, \pi]^{Lm}$.

Remark 2.3 We stress that our results strongly depends on choosing $\mathcal{O}$ as the sum of single-qubit observables over all the qubits. Indeed, if $\mathcal{O}$ is given by a single single-qubit observable, or more generally by an observable with eigenvalues ± 1, the training does not in general happen in the lazy regime [41].

[5] up to an irrelevant phase.

We are going to consider, as a model function, the *expected value* of the measurement of $\mathcal{M}$ on the initial state $|0^m\rangle \equiv \otimes_{k=1}^m |0\rangle_k \in \mathcal{H}$, up to a normalization $N(m)$:

$$\begin{aligned} f^{(m)}(\Theta, x) &= \frac{1}{N(m)} \langle 0^m | U^\dagger(\Theta, x) \mathcal{O} U(\Theta, x) | 0^m \rangle \\ &= \frac{1}{N(m)} \sum_{k=1}^{m} f_k^{(m)}(\Theta, x), \end{aligned} \tag{2.12}$$

where we defined

$$f_k^{(m)}(\Theta, x) = \langle 0^m | U^\dagger(\Theta, x) \mathcal{O}_k U(\Theta, x) | 0^m \rangle. \tag{2.13}$$

$N(m)$ is determined by the covariance function of the model at initialization, i.e., when the parameters Θ are randomly initialized before the training (see Assumption 3.6). A circuit suffering of the problem of barren plateaus would have a normalization $N(m)$ exponentially decreasing as a function of m.

Remark 2.4 Because of Assumption 2.2, $f_k^{(m)}(\Theta, x)$ are bounded functions:

$$|f_k^{(m)}(\Theta, x)| \leq 1. \tag{2.14}$$

We will use $f(\Theta, x)$ instead of $f^{(m)}(\Theta, x)$, and $f_k(\Theta, x)$ instead of $f_k^{(m)}(\Theta, x)$ in order to have a cleaner notation.

2.2.2 The encoding of the input

The internal structure of $V_\ell(x)$ can be arbitrary as long as the second point of Definition 2.2 is ensured, i.e., each qubit interacts at most with one different qubit. This general requirement will be enough for the discussion of the case with finite $\mathcal{X}$, so the following restriction is meaningful only for the reader interested in the general case of infinite $\mathcal{X}$ (section 6).

Let S_ℓ be the set of the couples $\{i, j\}$ of qubits i, j interacting in the layer ℓ. If a qubit i does not interact with any other qubit in the layer ℓ then we will just add the singlet $\{i\}$ to S_ℓ. Given this definition we will suppose that the unitaries $V(x)$ have the form

$$V_\ell(x) = \prod_{E \in S_\ell} \prod_{k=1}^{\dim \mathcal{X}} U_{E,k}^{(\ell)} \prod_{j \in E} e^{-i x_k \mathcal{K}_k^{\ell, j}} \tag{2.15}$$

where

1. $U_{E,k}^{(\ell)}$ is an arbitrary non parameterized unitary acting only on the qubits belonging to E;
2. $\mathcal{K}_k^{\ell,j}$ acts only on the qubit j;
3. Either $\mathrm{Spec}(\mathcal{K}_k^{\ell,j}) = \{-1, +1\}$ (in this case, the ℓ-th layer encodes x_k nontrivially in the qubit j) or $\mathcal{K}_k^{\ell,j} = 0$ (in this case, the ℓ-th layer does not encode x_k in the qubit j).

As a consequence of the third point, the model function $f(\Theta, \cdot)$ is componentwise periodic, with period π for any $\Theta \in \mathcal{X}$, so we will consider $\mathcal{X} = [0, \pi]^{\dim \mathcal{X}}$. In conclusion,

$$\begin{cases} U(\Theta, x) = V_L(x) W_L(\Theta) \cdots V_1(x) W_1(\Theta), \\ f(\Theta, x) = \dfrac{1}{N(m)} \langle 0^m | U^\dagger(\Theta, x) \mathcal{O} U(\Theta, x) | 0^m \rangle. \end{cases} \tag{2.16}$$

Remark 2.5 In Definition 2.2 we asked each qubit to be acted on *by at most one gate* per layer. Even if in (2.15) it might appear that more than one gate per qubit is acting, it is sufficient to notice that $V_\ell(x)$ can be rewritten in the form

$$V_\ell(x) = \bigotimes_{E \in S_\ell} U_E^{(\ell)}(x), \tag{2.17}$$

where $U_E^{(\ell)}(x)$ are single-qubit or two-qubit unitaries fixed by (2.15), to conclude that Definition 2.2 holds.

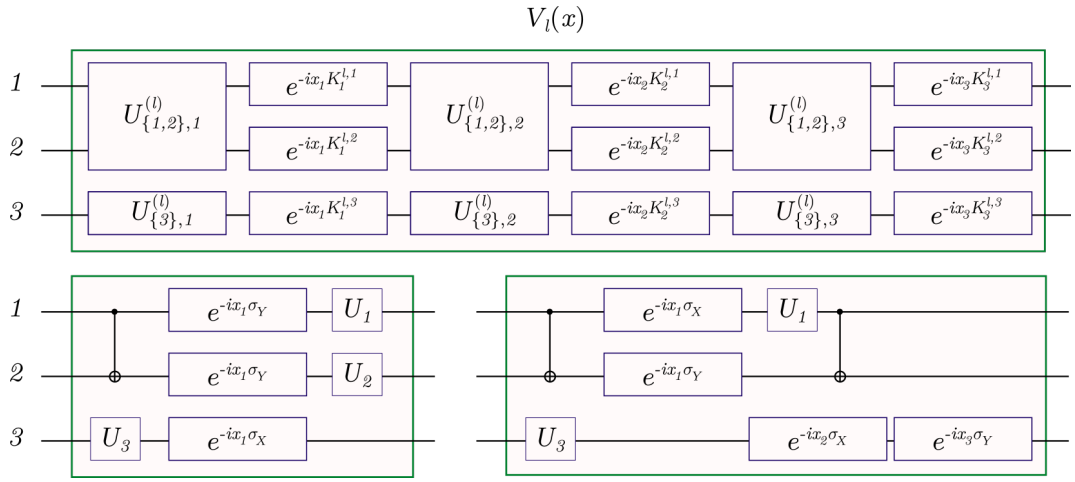


Fig. 4 The general structure of a feature encoding layer (above) for $m = 3$, $\dim \mathcal{X} = 3$ and $S_\ell = \{\{1,2\},\{3\}\}$ according to (2.15) with a couple of examples.

2.3 Light cones

2.3.1 Light cones and their extensions

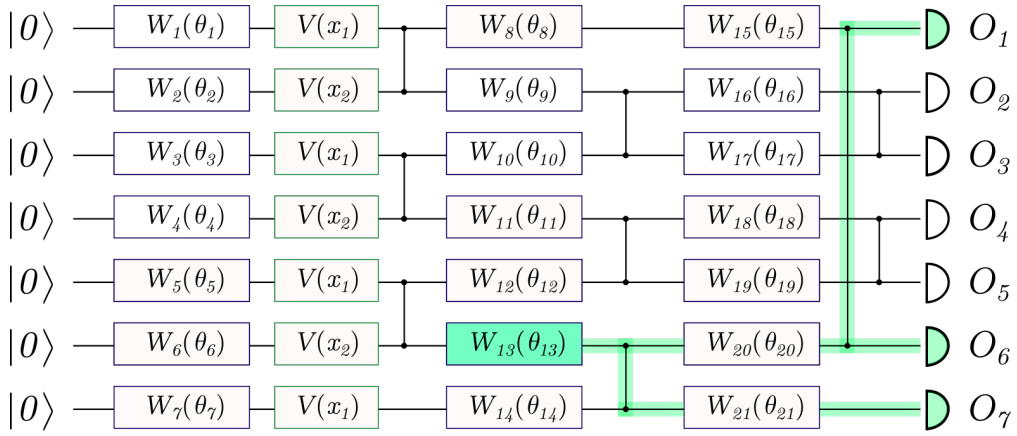


Fig. 5 Extended light cone $\mathcal{M}_{13} = \{1, 6, 7\}$ of the parameter θ_{13} for the circuit in the figure. Here $m = 7$, $\mathcal{X} = [0, \pi]^2$, $|\Theta| = 21$. Informally, the set $\mathcal{M}_{13}$ is the answer to the question: what are all the observables that may depend on the parameter θ_{13}?

Light cones are objects which, together with $N(m)$, characterize the *global* properties of the architecture of the circuit and will be crucial to understand the interplay between

1. conditions for trainability,
2. conditions for quantum advantages.

In this subsection we are going to introduce the light cone of a parameter and of a local observable, and we will define the local Hilbert space of an observable. We start from a simple definition, which is useful to understand the meaning of these sets (Definition 2.4). Then, we will give an operative definition to build some sets containing the light cones: we will call them *extended light cones* (Definition 2.5). Since the second definition allows cleaner proofs and upper bounds, we will use it in all the computations of the following sections.

Definition 2.4 (Light cones) For any parameter index $i \in \{1, \dots, |\Theta|\}$ we define the *future light cone* $\mathscr{L}_i^f$ *of the parameter* θ_i as the subset

$$\mathscr{L}_i^f = \{k \in \{1, \dots, m\} : f_k(\Theta, x) \text{ depends on } \theta_i\}. \tag{2.18}$$

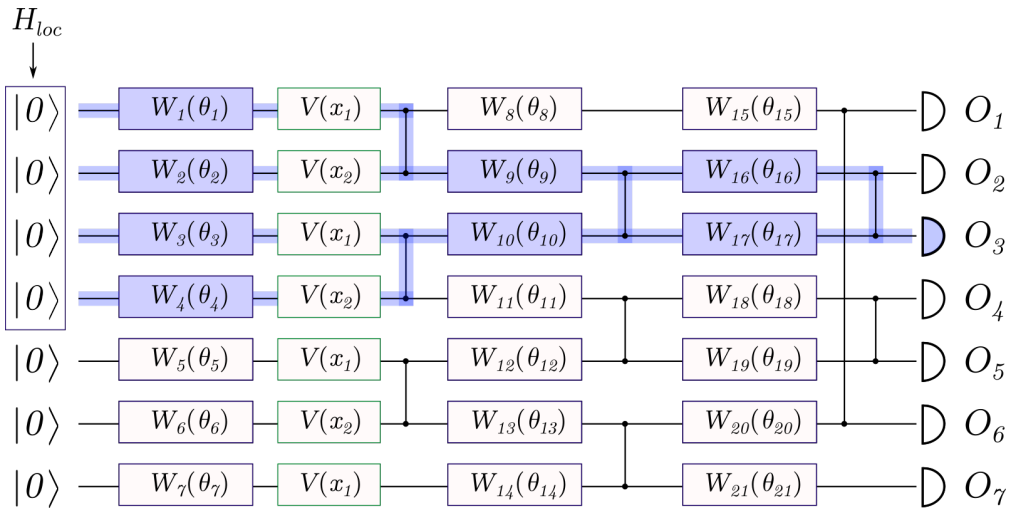


Fig. 6 Extended light cone $\mathcal{N}_3 = \{1, 2, 3, 4, 9, 10, 16, 17\}$ of the observable $f_3(\Theta, x)$ for the circuit in the figure. Here $m = 7$, $\mathcal{X} = [0, \pi]^2$, $|\Theta| = 21$. $\mathcal{H}_{\text{loc}}$ is the local Hilbert space associated to $f_3(\Theta, x)$. Informally, the set $\mathcal{N}_3$ is the answer to the question: what are all the parameters that may influence the observable $f_3(\Theta, x)$?

Besides, for any qubit index $k \in \{1, \dots, m\}$ we define the *past light cone* $\mathscr{L}_k^p$ *of the qubit* k as the subset

$$\mathscr{L}_k^p = \{i \in \{1, \dots, |\Theta|\} : f_k(\Theta, x) \text{ depends on } \theta_i\}. \tag{2.19}$$

Both $\mathscr{L}_i^f$ and $\mathscr{L}_k^p$ are useful to track the dependence of the observables on the parameters. More precisely, we want to understand which observables are affected by the random initialization of a parameter and by its evolution during the training. The following definition generalizes these sets (see Corollary 2.5), giving an operative way to build them by the only knowledge of the interactions between the qubits of the circuit.

For any circuit U, we need to define some auxiliary sets:

$$\mathcal{I}_{\ell,k} = \{k' \in \{1, \dots, m\} : \text{ the qubit } k \text{ interacts with the qubit } k' \text{ in the layer } \ell\} \cup \{k\}, \tag{2.20}$$

$$\mathcal{J}_k^\ell = \begin{cases} \mathcal{I}_{L,k} & \ell = L \\ \bigcup_{k' \in \mathcal{J}_k^{\ell+1}} \mathcal{I}_{\ell,k'} & \ell < L \,, \end{cases} \tag{2.21}$$

$$\mathcal{N}_k^\ell = \bigcup_{k' \in \mathcal{J}_k^\ell} \{[\ell\, k']\}. \tag{2.22}$$

In particular $\mathcal{J}_k^1$ is the set of qubits in the past light cone of the observable k, i.e., the qubits involved in the computation of its expectation value. The *family of interactions* $\{\mathcal{I}_{\ell,k}\}_{\ell=1,\dots,L;k=1,\dots,m}$ will be denoted as $\mathcal{I}_U$. When we will consider the sets (2.21) and (2.22) referred to a different circuit U', we will use the notation

$$\mathcal{J}_k^\ell\Big|_{U'}, \qquad \mathcal{N}_k^\ell\Big|_{U'}. \tag{2.23}$$

Now we have all the ingredients to define a generalization of the light cones.

Definition 2.5 (Extended light cones) Given any qubit index $k \in \{1, \dots, m\}$ we define the *extended past light cone* $\mathcal{N}_k$ *of the qubit* k as the subset of the parameter indices $\{1, \dots, |\Theta|\}$ given by

$$\mathcal{N}_k = \bigcup_{\ell=1}^{L} \mathcal{N}_k^\ell. \tag{2.24}$$

Furthermore, given any parameter index $i \in \{1, \dots, |\Theta|\}$, we define the *extended future light cone* $\mathcal{M}_i$ *of the parameter* θ_i as the subset of the qubit indices $\{1, \dots, m\}$ with the following property:

$$\mathcal{M}_i = \{k \in \{1, \dots, m\} : i \in \mathcal{N}_k\}. \tag{2.25}$$

Lemma 2.3 (Constructive definition of the extended future light cones) *As a consequence of Definition 2.5, we have a procedure to construct the extended future light cones of the parameters using the family of interactions $\mathcal{I}_U$:*

$$\begin{cases} \mathcal{M}_{[Lk]} = \mathcal{I}_{L,k} & \ell = L, \\ \mathcal{M}_{[\ell k]} = \bigcup\limits_{k' \in \mathcal{I}_{\ell,k}} \mathcal{M}_{[(\ell+1)\, k']} & \ell < L. \end{cases} \tag{2.26}$$

Proof Using (2.22),

$$[\ell k] \in \mathcal{N}_{k'} \quad \Longleftrightarrow \quad [\ell k] \in \mathcal{N}^{\ell}_{k'} \iff \quad k \in \mathcal{J}^{\ell}_{k'}, \tag{2.27}$$

and by Definition 2.5,

$$\mathcal{M}_{[\ell k]} = \{k' : [\ell k] \in \mathcal{N}_{k'}\} = \{k' : [\ell k] \in \mathcal{N}^{\ell}_{k'}\} = \{k' : k \in \mathcal{J}^{\ell}_{k'}\}. \tag{2.28}$$

By (2.21), if $\ell < L$,

$$k \in \mathcal{J}^{\ell}_{k'} \quad \Longleftrightarrow \quad k \in \mathcal{I}_{\ell,\bar{k}} \text{ for some } \bar{k} \in \mathcal{J}^{\ell+1}_{k'} \quad \Longleftrightarrow \quad \bar{k} \in \mathcal{I}_{\ell,k} \text{ for some } \bar{k} \in \mathcal{J}^{\ell+1}_{k'}, \tag{2.29}$$

so

$$\mathcal{M}_{[\ell k]} = \{k' : k \in \mathcal{J}^{\ell}_{k'}\} = \bigcup_{\bar{k} \in \mathcal{I}_{\ell,k}} \{k' : \bar{k} \in \mathcal{J}^{\ell+1}_{k'}\} = \bigcup_{\bar{k} \in \mathcal{I}_{\ell,k}} \mathcal{M}_{[(\ell+1)\, \bar{k}]}. \tag{2.30}$$

Finally,

$$\mathcal{M}_{[Lk]} = \{k' : k \in \mathcal{I}_{L,k'}\} = \{k' : k' \in \mathcal{I}_{L,k}\} = \mathcal{I}_{L,k}. \tag{2.31}$$

□

Definition 2.6 The *maximal cardinalities* of the extended light cones will be denoted as

$$|\mathcal{M}| = \max_i |\mathcal{M}_i|, \qquad |\mathcal{N}| = \max_k |\mathcal{N}_k| \tag{2.32}$$

and, for any integer $k \geq 1$, the sum of the k-th power of the cardinalities of $\mathcal{M}_i$ will be called

$$\Sigma_k = \sum_{i=1}^{|\Theta|} |\mathcal{M}_i|^k. \tag{2.33}$$

Remark 2.6 Clearly, the simplest upper bound to Σ_k in terms of $|\mathcal{M}|$ is

$$\Sigma_k \leq Lm|\mathcal{M}|^k. \tag{2.34}$$

2.3.2 The circuit seen by a local observable

The aim of this subsubsection is to determine the gates in the circuit on which the expected value of each local observable actually depends (see Figure 7). We start from the following remark.

Remark 2.7 According to Definition 2.2, we can rewrite

$$V_\ell(x) = \prod_{\{(k,k') : k' \in \mathcal{I}_{\ell,k}\}/\sim} V^{(k,k')}_\ell, \tag{2.35}$$

where $V^{(k,k')}_\ell$ is a unitary operation acting only on the qubit k and on the qubit k', and $\sim$ is the equivalence relation $(k,k') \sim (k',k)$. The order of the product of the unitaries $V^{(k,k')}_\ell$ is irrelevant, since they commute as they act on different Hilbert spaces.

Definition 2.7 (Pruning of a circuit) The *pruning* operation $[\,\cdot\,]_k$ $(k = 1, \ldots, m)$ of the circuit U to which the family $\mathcal{I}_U$ is associated,

$$\big[U\big]_k = \big[U_L U_{L-1} \ldots U_1\big]_k = \left[\prod_{\ell=0}^{L} V_\ell(x) W_\ell(\Theta)\right]_k, \tag{2.36}$$

is the circuit obtained from U by replacing, for any $k', k'' \in \{1, \ldots, m\}$ and $\ell \in \{1, \ldots, L\}$,

$$W_{[\ell k']} \quad \to \quad \mathbb{1} \qquad \Longleftrightarrow \qquad [\ell k'] \notin \mathcal{N}^{\ell}_k, \tag{2.37}$$

$$V^{(k',k'')}_\ell \quad \to \quad \mathbb{1} \qquad \Longleftrightarrow \qquad k'' \in \mathcal{I}_{\ell,k'} \text{ and } \{k', k''\} \cap \mathcal{J}^{\ell}_k = \emptyset. \tag{2.38}$$

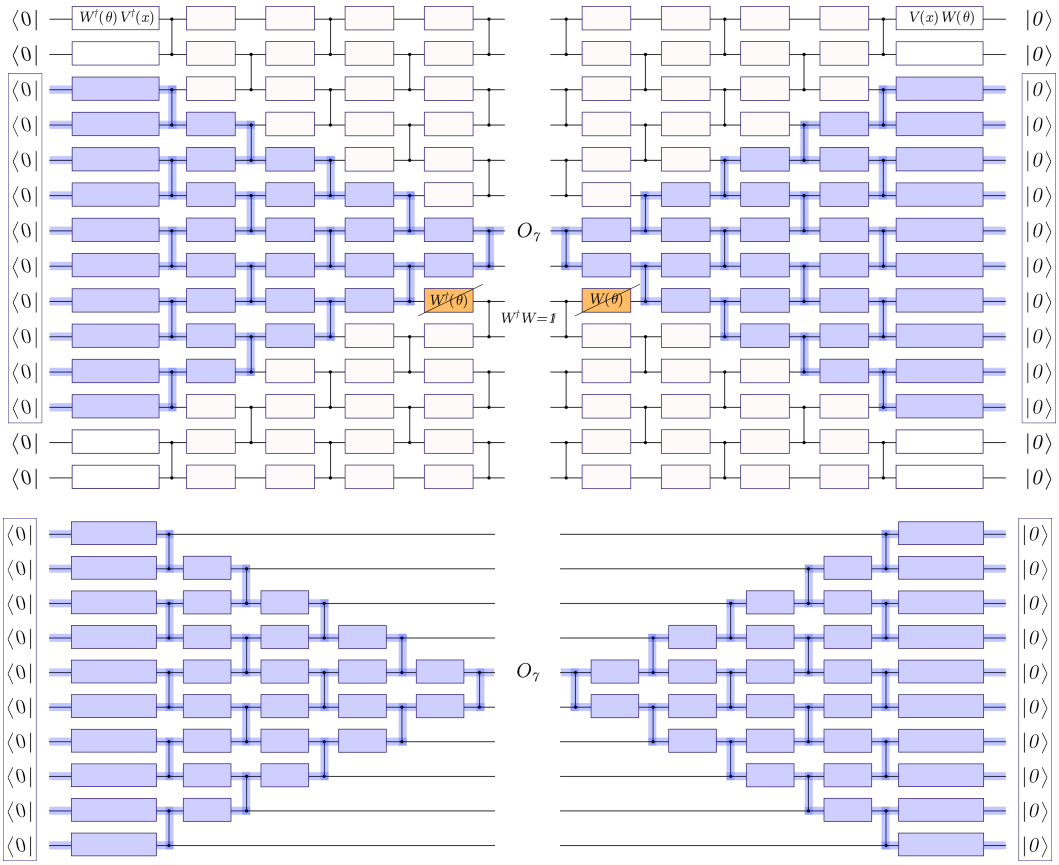


Fig. 7 Above: the entire circuit. Below: the part of the circuit which contributes to the computation of $\langle 0^m|U^\dagger(\Theta,x)\mathcal{O}_7 U(\Theta,x)|0^m\rangle$. We will call it the "pruned" circuit.

Lemma 2.4 (Fundamental property of the pruning operation) *Given any circuit $U(\Theta,x)$ according to Definition 2.2 and a local observable $\mathcal{O}_k$ acting only on the qubit k, for any $|\psi\rangle \in \mathcal{H}$ the following identity holds*

$$\langle\psi|U^\dagger(\Theta,x)\mathcal{O}_k U(\Theta,x)|\psi\rangle = \langle\psi|\big[U(\Theta,x)\big]_k^\dagger \mathcal{O}_k \big[U(\Theta,x)\big]_k|\psi\rangle. \tag{2.39}$$

Proof We prove (2.39) by induction on L. If $L=0$, $U=\mathbb{1}$ and $[U]_k=\mathbb{1}$, the claim holds. Let us now consider a circuit composed of $L>0$ layers

$$U = U_L U_{L-1}\dots U_2 U_1 \tag{2.40}$$

and assume the claim for $L-1$ layers. The unitary

$$U' = U_L U_{L-1}\dots U_2 \tag{2.41}$$

represents a circuit of $L-1$ layers. Since the sets $\mathcal{J}_k^\ell$ and $\mathcal{N}_k^\ell$ are defined starting from the last layer, the circuits U and U' have in common

$$\mathcal{J}_k^\ell\Big|_U = \mathcal{J}_k^{\ell-1}\Big|_{U'} \quad \text{and} \quad \mathcal{N}_k^\ell\Big|_U = \mathcal{N}_k^{\ell-1}\Big|_{U'} \qquad \forall\,\ell>1 \tag{2.42}$$

Let $\big|\psi'\big\rangle = U_1\,|\psi\rangle$; by inductive hypothesis

$$\begin{aligned}\langle\psi'|(U')^\dagger\mathcal{O}_k U'|\psi'\rangle &= \langle\psi'|\big[U'\big]_k^\dagger\mathcal{O}_k\big[U'\big]_k|\psi'\rangle\\ &= \langle\psi|U_1^\dagger\big[U'\big]_k^\dagger\mathcal{O}_k\big[U'\big]_k U_1|\psi\rangle\\ &= \langle\psi|\,(V_1W_1)^\dagger\,\big[U'\big]_k^\dagger\mathcal{O}_k\big[U'\big]_k\,(V_1W_1)\,|\psi\rangle.\end{aligned} \tag{2.43}$$

Suppose $\{k',k''\}\cap\mathcal{J}_k^\ell=\emptyset$ where $k''\in\mathcal{I}_{1,k'}$; then

$$V_1^{(k',k'')} \quad \text{commutes with} \quad \big[U'\big]_k \tag{2.44}$$

since $\left[U'\right]_k$ acts as the identity on the qubits k', k'', due to the definition of the pruning operation combined with the elementary property $\mathcal{J}_k^\ell \subseteq \mathcal{J}_k^{\ell+1}$ for any $\ell = 1, \dots, L-1$. Furthermore, because $k \in \mathcal{J}_k^\ell$, both k' and k'' are different from k, whence

$$\left(V_1^{(k',k'')}\right)^\dagger \mathcal{O}_k V_1^{(k',k'')} = \mathcal{O}_k \left(V_1^{(k',k'')}\right)^\dagger V_1^{(k',k'')} = \mathcal{O}_k. \tag{2.45}$$

Similarly, if $[\ell k'] \notin \mathcal{N}_k^\ell$, then

$$W_{[1k']} \quad \text{commutes with} \quad \left[U'\right]_k \quad \text{and} \quad \mathcal{O}_k. \tag{2.46}$$

As a consequence,

$$\begin{aligned} \langle\psi'|(U')^\dagger \mathcal{O}_k U'|\psi'\rangle &= \langle\psi| \left(V_1 W_1\right)^\dagger \left[U'\right]_k^\dagger \mathcal{O}_k \left[U'\right]_k \left(V_1 W_1\right)|\psi\rangle \\ &= \langle\psi| \left[U' V_1 W_1\right]_k^\dagger \mathcal{O}_k \left[U' V_1 W_1\right]_k |\psi\rangle \\ &= \langle\psi| \left[U\right]_k^\dagger \mathcal{O}_k \left[U\right]_k |\psi\rangle, \end{aligned} \tag{2.47}$$

which proves (2.39) for L layers. □

As a corollary, we can show the property anticipated above.

Corollary 2.5 (Extended light cones generalize the light cones) *The following relations hold*

$$\mathscr{L}_k^p \subseteq \mathcal{N}_k, \qquad \mathscr{L}_i^f \subseteq \mathcal{M}_i. \tag{2.48}$$

Proof If $f_k(\Theta, x)$ depends on $\theta_{[\ell k']}$, then, by Lemma 2.4 the parameter must appear also in the pruned circuit

$$\left[U(\Theta, x)\right]_k. \tag{2.49}$$

By definition of pruning, the dependence on the parameter $\theta_{[\ell k']}$ is not removed if and only if $[\ell k'] \in \mathcal{N}_k^\ell$. Since

$$\mathcal{N}_k = \bigcup_{\ell=1}^{L} \mathcal{N}_k^\ell \tag{2.50}$$

then $\theta_{[\ell k']} \in \mathcal{N}_k$, whence

$$\mathcal{N}_k \supseteq \mathscr{L}_k^p. \tag{2.51}$$

This implies that

$$\mathscr{L}_i^f = \{k : i \in \mathscr{L}_k^p\} \subseteq \{k : i \in \mathcal{N}_k\} = \mathcal{M}_i. \tag{2.52}$$

□

Remark 2.8 Since the extended light cones are immediate to construct[6] using Definition 2.5, from now on we will consider only them and, with a slight abuse of notation, we will simply refer to them as "light cones".

In the following sections, when we will need to emphasize the dependence of a local observable $f_k(\Theta, x)$ on the subset of parameter $\{\theta_i : i \in \mathcal{N}_k\}$, we will use the notation (borrowed from [38]) $f_k(\Theta_{\mathcal{N}_k}, x)$ instead of $f_k(\Theta, x)$. More generally, if we will need to underline the dependence of a function $g(\Theta)$ on a subset $S \subseteq \{\theta_1, \dots, \theta_{Lm}\}$, we will write $g(\Theta_S)$ instead of $g(\Theta)$.

2.3.3 Light cones, local Hilbert spaces and classical simulability

As we can clearly notice by looking at Figure 7 and as Theorem 2.4 implies, the Hilbert space involved in the computations of the output $f_k(\Theta, x)$ of a local observable is smaller that $\mathcal{H}$. This motivates the following definition.

Definition 2.8 (Local Hilbert space) If we denote as $\mathcal{H}_1, \mathcal{H}_2, \dots, \mathcal{H}_m$ the Hilbert space associated to the qubits $1, 2, \dots, m$, the *local Hilbert space* associated with a local observable $\mathcal{O}_k$ is defined as

$$\mathcal{H}_{\text{loc}}^k = \bigotimes_{k' \in \mathcal{J}_k^1} \mathcal{H}_{k'}. \tag{2.53}$$

[6] Instead, the original light cones may be more difficult to construct, when some cancellations of the dependencies occurs or when $\mathcal{G}_i = 0$ for some i. The knowledege of the family $\mathcal{I}_U$ is insufficient to define $\mathscr{L}_k^p$ and $\mathscr{L}_i^f$, while is the only object involved in the construction of $\mathcal{N}_k$ and $q\mathcal{M}_i$.

In particular

$$\dim \mathcal{H}^k_{\rm loc} = 2^{|\mathcal{J}^1_k|}. \tag{2.54}$$

Lemma 2.6 (Sufficiency of the local Hilbert space) *The computation of $f_k(\Theta, x)$ requires only linear-algebra operations in $\mathcal{H}^k_{\rm loc}$.*

Proof Because of Theorem 2.4, the only qubits involved in the computation of $f_k(\Theta, x)$ are the ones on which $[U]_k$ acts nontrivially. In particular, $[U]_k$ is the identity outside the tensor product of the Hilbert spaces of the qubits

$$\bigcup_{\ell=1}^{L} \mathcal{I}^\ell_k = \mathcal{I}^1_k, \tag{2.55}$$

where we used that $\mathcal{I}^{\ell+1}_k \subseteq \mathcal{I}^\ell_k$ for any $\ell = 1, \dots, L-1$. Such tensor product is (2.53). □

A necessary condition to achieve a quantum advantage with respect to classical computations is the requirement that the dynamics of a quantum circuit must be hard to simulate classically. The linear-algebra operations required to compute the model function $f(\Theta, x)$ are of the order of m times the operations required to compute a local observable $f_k(\Theta, x)$. The order of the number of these operations is $O(\text{poly}(\dim \mathcal{H}_{\rm loc}))$.

Lemma 2.7 (Necessary condition for quantum advantages) *We can estimate the dimension of the local Hilbert space:*

$$2^{|\mathcal{N}|/L} \leq \max_k \dim \mathcal{H}^k_{\rm loc} \leq 2^{|\mathcal{N}|} \tag{2.56}$$

Therefore, a necessary condition to achieve a quantum advantage is that $|\mathcal{N}|$ grows superlogarithmically with the number of qubits.

Any logarithmic or sublogarithmic growth of $|\mathcal{N}|$ excludes any superpolynomial quantum advantage.

Proof By Definition 2.5, and using the fact that $\mathcal{N}^1_k \supseteq \mathcal{N}^\ell_K$ for any $\ell = 1, \dots, L$,

$$\mathcal{N}_k = \bigcup_{\ell=1}^{L} \mathcal{N}^\ell_k \qquad \to \qquad |\mathcal{N}_k| \leq L|\mathcal{N}^1_k| \quad \text{and} \quad |\mathcal{N}^1_k| \leq |\mathcal{N}|, \tag{2.57}$$

Since $|\mathcal{N}^1_k| = \log_2 \dim \mathcal{H}^k_{\rm loc}$,

$$\frac{|\mathcal{N}_k|}{L} \leq \log_2 \dim \mathcal{H}^k_{\rm loc} \leq |\mathcal{N}|. \tag{2.58}$$

Taking the maximum over k,

$$\frac{|\mathcal{N}|}{L} \leq \max_k \log_2 \dim \mathcal{H}^k_{\rm loc} \leq |\mathcal{N}|. \tag{2.59}$$

□

Remark 2.9 The fact that the outcome of any local observable only depends on the state reduced to the local Hilbert space has operational consequences when the circuit has to be experimentally realized and gives further insights about the meaning of the infinite width limit in quantum neural networks.

1. Concerning the implementation, wide architectures do not actually require a quantum computer with as many qubits as their widths: the number of qubits which are needed in the laboratory is not m, but $\max_k \log \dim \mathcal{H}^k_{\rm loc} = \max_k |\mathcal{J}^1_k|$ (see (2.54)), since the expectation value of each measured one-qubit observable can be estimated independently, and such estimate requires only the qubits in the past light-cone of the measured qubit. Without producing architectures which can be naively classically simulated, it is possible to choose $|\mathcal{J}^1_k| = O((\epsilon \log m)^d)$ with $d \geq 2$ and $\epsilon > 0$ smaller than a certain threshold, as we will discuss in detail in subsection 2.5. This means that the effective number of qubits needed to run a quantum neural network of size m is much smaller than m.
2. Since number of qubits involved in the light cones is the key determining the effective size of the implemented network and the classical (non-)simulability properties, we should think about $\max_k |\mathcal{J}^1_k|$ as the effective width of the circuit.

2.3.4 Architecture-independent bounds and geometrically local circuits

Is it possible to estimate the cardinalities of the light cones without knowing the specific structure of the circuit? Yes, it is, but the bounds are very weak:

$$|\mathcal{M}| \leq 2^L \qquad |\mathcal{N}| \leq 2^{L+1} \qquad \Sigma_n \leq 2m2^{nL} \tag{2.60}$$

For the derivation of these bounds, see Appendix B.

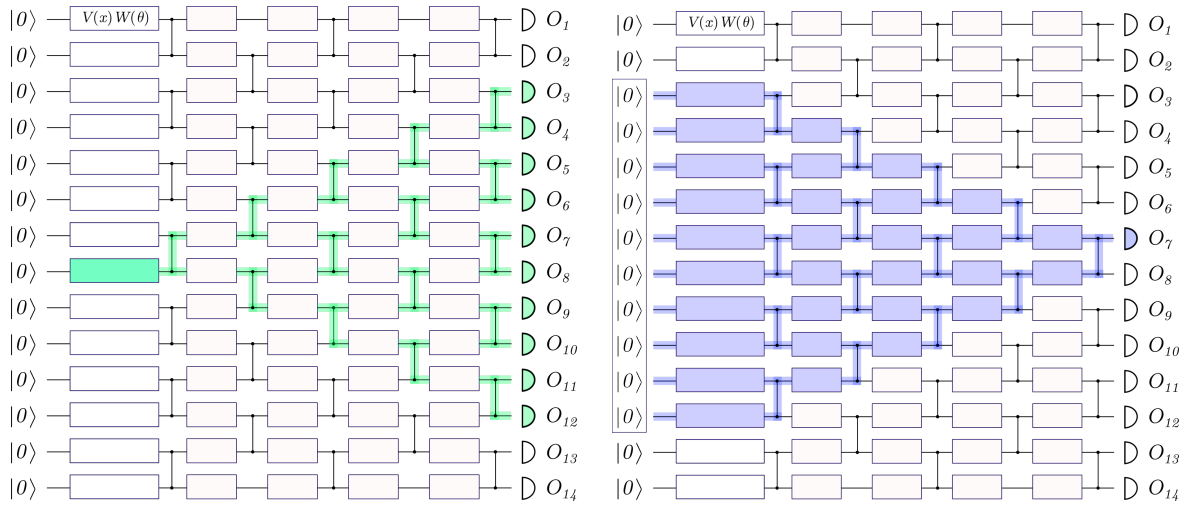


Fig. 8 Light cones $\mathcal{M}_{[1\,8]}$ and $\mathcal{N}_7$ for a geometrically local circuit. Since each qubit interacts only with neighbouring qubits, the growth of the light cones is polynomial in the number of layers L.

All the bounds obtained for the most generic architecture using 2-qubit gates are exponential in L, which in general is not a good dependence if we are interested in the case $L = L(m)$. These estimates – which accounts also for the worst case – are typically too loose since, in a physical implementation with many qubits, the reasonable assumption that only neighbouring qubits interact[7] turns this exponential behaviour into a polynomial growth. This can be informally realized by looking at Figure 8. In particular, the dependence on the number of qubits is of the order

$$\dim \mathcal{H}_{\text{loc}} = 2^{O(L^d)}, \qquad |\mathcal{N}| = O(L \cdot L^d), \qquad |\mathcal{M}| = O(L^d), \tag{2.61}$$

where d is the dimension of the lattice on which qubits are arranged. We will not enter into the details of these circuits since in our theorems we never assume geometrical locality, and all the theorems will be stated in terms of the maximal cardinalities of the light cones.

We will come back to the problem of the exponential dependence on the number of layers in subsection 2.5.

2.4 Random initialization of quantum circuits

We say that a circuit is randomly initialized when the parameters θ_i are sampled from independent (not necessarily identical) distributions. At the moment, we do not add any assumption on these distributions other than the independence and the following requirement.

Assumption 2.8 *The final layer is chosen so that, for any* $k = 1, \ldots, m$,

$$\mathbb{E}[f_k(\Theta, x)] = 0. \tag{2.62}$$

[7] Such architectures are called *geometrically local* [38].

Now we show that any circuit $U(\Theta, x)$ with traceless local observables satisfies (2.62) by means of a specific final layer of local parameterized gates. To understand how this works, let us consider the simpler case in Figure 9 where the observables are Pauli Z.

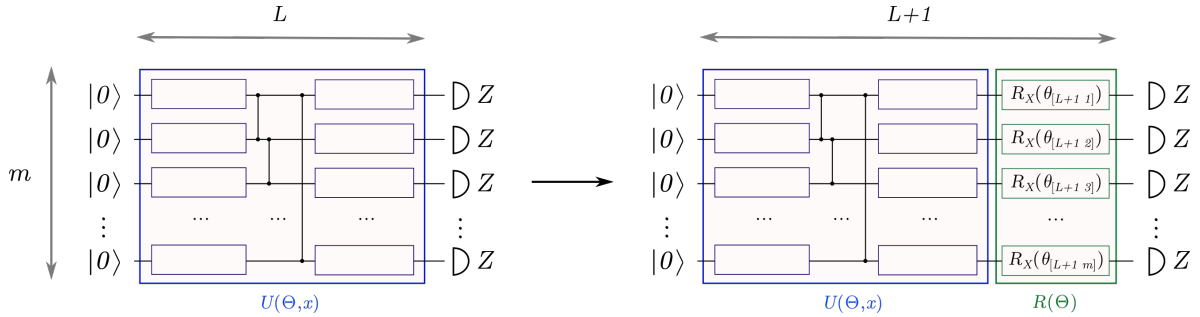


Fig. 9 The zero mean hypothesis can be obtained by adding a final layer.

Let us add a final layer

$$U_{L+1}(\Theta) = R_X(\theta_{[(L+1)\,1]}) \otimes R_X(\theta_{[(L+1)\,2]}) \otimes \cdots \otimes R_X(\theta_{[(L+1)\,m]}), \tag{2.63}$$

where

$$R_X(\theta) = e^{-i\theta\sigma_X} \qquad \sigma_X = \begin{pmatrix} & 1 \\ 1 & \end{pmatrix}. \tag{2.64}$$

Therefore

$$\begin{aligned} f'_k(\Theta, x) = \langle 0^m | \, U^\dagger(\Theta, x) U^\dagger_{L+1}(\Theta) \\ \mathbb{1}_1 \otimes \cdots \otimes \mathbb{1}_{k-1} \otimes Z_k \otimes \mathbb{1}_{k+1} \otimes \cdots \otimes \mathbb{1}_m \\ U_{L+1}(\Theta) U(\Theta, x) \, |0^m\rangle \\ = \langle 0^m | \, U^\dagger(\Theta, x) \\ \mathbb{1}_1 \otimes \cdots \otimes e^{i\theta_{[(L+1)\,k]}\sigma_X} Z_k e^{-i\theta_{[(L+1)\,k]}\sigma_X} \otimes \cdots \otimes \mathbb{1}_m \\ U(\Theta, x) \, |0^m\rangle \, , \end{aligned} \tag{2.65}$$

where X_k and Z_k are respectively σ_X and σ_Z acting on the qubit k. Using the fact that

$$\sigma_i \sigma_j = \delta_{ij} + i\epsilon_{ijk}\sigma_k \tag{2.66}$$

and

$$e^{-i\theta\hat{n}\cdot\vec{\sigma}} = \mathbb{1}\cos\theta - i\hat{n}\cdot\vec{\sigma}\sin\theta \quad \text{with} \quad |\hat{n}| = 1, \tag{2.67}$$

we see that

$$\begin{aligned} e^{i\theta\sigma_X}\sigma_Z e^{-i\theta\sigma_X} &= (\mathbb{1}\cos\theta + i\sigma_X\sin\theta)\sigma_Z(\mathbb{1}\cos\theta - i\sigma_X\sin\theta) \\ &= (\mathbb{1}\cos\theta + i\sigma_X\sin\theta)(\sigma_Z\cos\theta + \sigma_Y\sin\theta) \\ &= \sigma_Z\cos^2\theta + \sigma_Y\cos\theta\sin\theta + \sigma_Y\sin\theta\cos\theta - \sigma_Z\sin^2\theta \\ &= \sigma_Z(\cos^2\theta - \sin^2\theta) + \sigma_Y 2\sin\theta\cos\theta \\ &= \sigma_Z\cos 2\theta + \sigma_Y\sin 2\theta. \end{aligned} \tag{2.68}$$

If θ is uniformly distributed in $[0, 2\pi]$, then

$$\mathbb{E}[\cos 2\theta] = 0 \qquad \text{and} \qquad \mathbb{E}[\sin 2\theta] = 0. \tag{2.69}$$

So, if, for all $k = 1, \ldots, m$, $\theta_{[(L+1)\,k]}$ is uniform in $[0, 2\pi]$ and independent of $\theta_1, \ldots, \theta_{Lm}$, then

$$f'_k(\Theta, x) = \langle 0^m | \, U^\dagger(\Theta, x) \mathbb{1}_{1,\ldots,k-1} \otimes Z_k \cos 2\theta_{[(L+1)\,k]} \otimes \mathbb{1}_{k+1,\ldots,m} U(\Theta, x) \, |0^m\rangle$$

$$
\begin{aligned}
&+ \langle 0^m | \, U^\dagger(\Theta, x) \mathbb{1}_{1,\dots,k-1} \otimes Y_k \sin 2\theta_{[(L+1)\,k]} \otimes \mathbb{1}_{k+1,\dots,m} U(\Theta, x) \, |0^m\rangle \\
&= f_k(\Theta, x) \cos 2\theta_{[(L+1)\,k]} + g_k(\Theta, x) \sin 2\theta_{[(L+1)\,k]};
\end{aligned} \tag{2.70}
$$

$$
\mathbb{E}[f'_k(\Theta, x)] = \mathbb{E}[f_k(\Theta, x)]\mathbb{E}[\cos 2\theta_{[(L+1)\,k]}] + \mathbb{E}[g_k(\Theta, x)]\mathbb{E}[\sin 2\theta_{[(L+1)\,k]}] = 0, \tag{2.71}
$$

for all $k \in \{1, \dots, m\}$ where

$$
\begin{aligned}
f_k(\Theta, x) &= \langle 0^m | U^\dagger(\Theta, x) \, (\mathbb{1}_{1,\dots,k-1} \otimes Z_k \otimes \mathbb{1}_{k+1,\dots,m}) \, U(\Theta, x) | 0^m \rangle \\
g_k(\Theta, x) &= \langle 0^m | U^\dagger(\Theta, x) \, (\mathbb{1}_{1,\dots,k-1} \otimes Y_k \otimes \mathbb{1}_{k+1,\dots,m}) \, U(\Theta, x) | 0^m \rangle
\end{aligned} \tag{2.72}
$$

is the output of the original circuit, which does not depend on $\theta_{[(L+1)\,k]}$. The general case is analogous.

The following lemma will be crucial to prove – in terms of the cardinalities of the light cones – that randomly initialized quantum neural networks yield model functions distributed as Gaussian processes.

Lemma 2.9 (Dependent observables at random initialization) *For any $k = 1, \dots, m$, let*

$$
\begin{aligned}
\mathcal{P}_k = \{k' \in \{1, \dots, m\} : f_{k'}(\Theta, x) \text{ is not independent of} \\
f_k(\Theta, x) \text{ at random initialization}\}.
\end{aligned} \tag{2.73}
$$

Then

$$
|\mathcal{P}_k| \leq |\mathcal{M}||\mathcal{N}|. \tag{2.74}
$$

Proof If the observables $f_k(\Theta, x)$ and $f_{k'}(\Theta, x)$ are not independent, then there is a parameter on which both the qubit k and the qubit k' depend:

$$
\begin{aligned}
\mathcal{P}_k &\subseteq \{k' \in \{1, \dots, m\} : \mathcal{N}_k \cap \mathcal{N}_{k'} \neq \emptyset\} \\
&= \{k' \in \{1, \dots, m\} : \exists i \in \mathcal{N}_k \cap \mathcal{N}_{k'}\} \\
&= \bigcup_{i \in \mathcal{N}_k} \{k' \in \{1, \dots, m\} : i \in \mathcal{N}_{k'}\} \\
&= \bigcup_{i \in \mathcal{N}_k} \{k' \in \{1, \dots, m\} : k' \in \mathcal{M}_i\} \\
&= \bigcup_{i \in \mathcal{N}_k} \mathcal{M}_i.
\end{aligned} \tag{2.75}
$$

Therefore

$$
|\mathcal{P}_k| \leq |\mathcal{N}_k| \max_i |\mathcal{M}_i| \leq |\mathcal{N}||\mathcal{M}|. \tag{2.76}
$$

□

2.5 Some examples and the estimate of $N(m)$

In the theorems of convergence and of trainability that we are going to prove in the next sections, the hypotheses will have the form

$$
\lim_{m \to \infty} \frac{L^\alpha m^\beta |\mathcal{M}|^\gamma |\mathcal{N}|^\delta}{N(m)} = 0 \tag{2.77}
$$

for some $\alpha, \beta, \gamma, \delta > 0$ according to the precise statement. It is important to provide concrete examples of architectures for which it is possible to estimate $N(m), |\mathcal{M}|$ and $|\mathcal{N}|$ so that the existence of valid candidates of circuits is ensured.

The example provided by E. Abedi & al. in [38] (see Figure 10) is a circuit in which

$$
N(m) = \sqrt{m}, \quad |\mathcal{M}| = O(L), \quad |\mathcal{N}| = O(L^2), \quad L \text{ is fixed.} \tag{2.78}
$$

As we will see, in general $\beta < \frac{1}{2}$, therefore this kind of circuit always satisfies all our theorems. However, since $\dim \mathcal{H}_{\text{loc}} = 2^{O(L)} = O(1)$, such circuit is classically simulable, so we do not expect to achieve

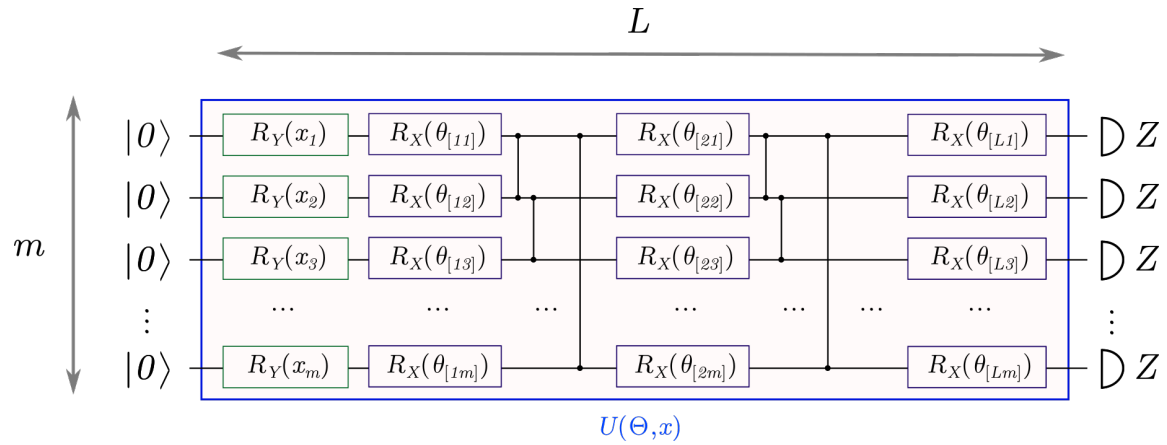


Fig. 10 The example described in [38].

quantum advantages with such architecture.
A suitable class of circuits for investigating quantum advantages is the one studied by J. C. Napp in [16], where the following bound is provided

$$N(m) \geq \frac{\sqrt{m}}{2^{CL}} \quad \text{for some fixed } C. \tag{2.79}$$

In a generic setting, the number of qubits in the past light cone of any observable O_k can grow as

$$|\mathcal{J}_k^1| = O(2^L). \tag{2.80}$$

Under the hypothesis of geometrical locality (i.e., each qubit can interact only with the nearest neighbour qubits), a d-dimensional lattice of qubits has

$$|\mathcal{J}_k^1| = O(L^d). \tag{2.81}$$

Let us assume to choose $L = \epsilon \log_2 m$. Without assumptions on the geometrical locality, $|\mathcal{J}_k^1| = O(m^\epsilon)$, which is an upper bound, so any growth $|\mathcal{J}_k^1| = \Theta(m^{\epsilon'})$ with $\epsilon' \leq \epsilon$ can be achieved by an appropriate choice of the interactions. In this case, each condition of the form (2.77) can be satisfied for ϵ, ϵ' small enough

$$\lim_{m\to\infty} \frac{(\log_2 m)^\alpha m^{\beta+(\gamma+\delta)\epsilon'}}{m^{1/2-C\epsilon}} = 0 \tag{2.82}$$

since in all our results the hypothesis appears with $\beta < 1/2$. However, the local Hilbert spaces have a superpolynomial dimension $2^{m^{\epsilon'}}$. In the geometrically local setting, we can choose $|\mathcal{J}_k^1| = \Theta((\epsilon \log_2 m)^d)$; if $d \geq 2$, then the local Hilbert spaces still have a superpolynomial dimension, since $2^{(\epsilon \log_2 m)^d} = m^{\epsilon^d \log_2^{d-1} m}$. Besides, each condition of the form (2.77) can be satisfied for ϵ small enough.

We should mention that the exponential decrease of $N(m)$ on the number of layers is a manifestation of the phenomenon of *barren plateaus*, which will be discussed in subsubsection 4.1.2. We will not further investigate the particular architectures which satisfy our theorems, but our statements will be motivated by the existence of examples such as the ones we have just presented.

As a final remark, we explain the deep reason to avoid in general the further simplification given by Lemma B.1. This lemma would reduce the hypothesis to the simpler (but stronger) form

$$\lim_{m\to\infty} \frac{L^\alpha m^\beta 2^{(\gamma+\delta)L}}{N(m)} = 0 \tag{2.83}$$

which requires only the knowledge of $N(m)$ and L. The exponential dependence on L (combined with the behaviour $N(m) = O(\sqrt{m})$, which in general is not tight) constrains L to be at most logarithmically dependent on m. The same bottleneck appears in the case of barren plateaus. Even though this constraint

does not impede a priori the achievement quantum advantages, such limit on the growth of L with respect to m might not be intrinsic: in particular, in the case of specific architectures avoiding barren plateaus, the simplification of Lemma B.1 would fictitiously rule out more interesting behaviours of $L(m)$. This is the reason why it is important to keep the explicit dependence on the maximal cardinalities of the light cones in the hypothesis of the theorems proved in the following sections.

3 Quantum neural networks with random parameters are Gaussian processes

In this section we are going to consider the random initialization of the parameters of the model function: $\{\theta_i\}_{i=1,\dots,Lm}$, are therefore random variables, which we assume to be independent and identically distributed. The goal of this section is to study the conditions under which the function generated by the quantum neural network is distributed as a Gaussian process as in the classical case (Theorem 3.7) and to fix reasonable assumptions on the output of the randomly initialized circuits (see Assumption 3.6).

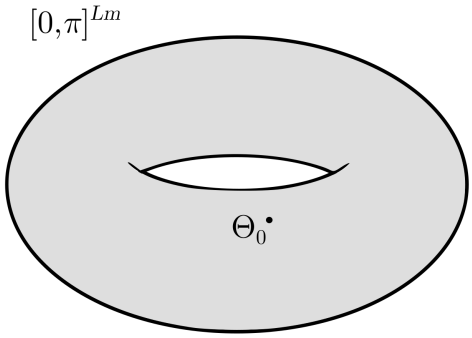


Fig. 11 The uniform random initialization in the parameter space $\mathscr{P} = [0,\pi]^{Lm}$.

As discussed in Remark 2.2, a typical parameter space could be $\mathscr{P} = [0,\pi]^{Lm}$ with periodic boundary conditions – which is a torus. Besides, a typical initialization could be the uniform one (with respect to Lebesgue measure on the subset $[0,\pi]^{Lm} \subseteq \mathbb{R}^{Lm}$). These particular choices are a convenient example to fix ideas on a possible initialization, but the assumptions and the theorem we will state in this section will be more general.

Differently from the classical neural networks [21], we cannot follow an inductive procedure by performing the limit $m \to \infty$ layer by layer: in our case, the number of qubits is the same in all the layers. Moreover, as discussed in subsubsection 2.3.3, the number of layer must grow with the number of qubits when we look for quantum advantages. Hence, if we proved that the output is distributed as a Gaussian process for any fixed L, this would be insufficient: the limit $L \to \infty$ could be performed only after the limit $m \to \infty$, but the dependence $L(m)$ requires a unique limit $m \to \infty$ in both the spatial (number of qubits) and temporal (number of layers) directions.

Our approach to face these issues is to consider the global properties of the random outputs and, in particular, their dependency relations. On the one hand, the strategy and the physical intuition behind the central limit theorem are important (see Theorem 3.2). On the other hand, looking back at the discussion about light cones and dependence between the local measurements, it is easy to realize that the model function is a sum of functions which, in general, are not independent. However, let us fix $k \in \{1,\dots,m\}$ and consider the observable $f_k(\Theta, x)$ after a random initialization of the parameters, which are *independent* random variables. For narrow enough light cones, the number of observables $f_{k'}(\Theta, x)$ whose output is not independent of the outcome of $f_k(\Theta, x)$ will be slowly growing with respect to the total number of qubits. This situation is depicted in Figure 13. Such behaviour can be thought as an approximate version of the hypothesis of the central limit theorem. If the deviations from the standard hypothesis are somehow "small", the result will still be valid.

Our results improve and generalize the discussion of [38] by means of quantitative bounds which are valid for any dependence $L(m)$ and which will be used to determine the behaviour of the circuit in the limit $m \to \infty$.

The outline of this section is the following: first, we will introduce the basic notions about the Gaussian processes and the theory of cumulants, and we will state some technical theorems from the literature to support quantitatively the intuition just presented (see subsection 3.1). Then, we will fix and discuss some assumptions about the output of the circuit (see subsection 3.2). Combining the technical results with these assumptions, we will prove the main theorem of the section (Theorem 3.7): a sufficient condition of the form (2.77) to ensure the convergence to the output to a Gaussian process. Finally, we will show, by means of a counterexample, that a circuit violating the hypothesis we will have just stated does not produce, in general, an output distributed as a Gaussian process (see Appendix A). Even though we expect that out hypotheses can be relaxed and the statement of the convergence theorem can be extended to a larger class of circuits using a different strategy for the proof, this observation excludes a generalization to any variational circuit.

3.1 Central limit theorem and generalizations

In this subsection we recall some fundamental definitions and results which serve as an introduction to our results and strategies. We start defining a Gaussian process (see [44] for more details).

3.1.1 Gaussian processes and the central limit theorem

Definition 3.1 (Stochastic process) A *stochastic process* is a collection of random variables $\{X_\alpha\}_{\alpha \in A}$, which defined on a common probability space.

When two random variables X and Y have the same distribution, we write $X \sim Y$.
Given $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, we write $X \sim \mathcal{N}(\mu, \sigma^2)$ when X is a random variable with the following Gaussian probability density function:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}. \tag{3.1}$$

If $\mu \in \mathbb{R}^N$ is a vector and $\mathcal{K} \in \mathbb{R}^{N \times N}$ is a positive-definite matrix, we write $X \sim \mathcal{N}(\mu, \mathcal{K})$ when X is a random vector of $\mathbb{R}^N$ with the following multivariate Gaussian probability density function:

$$p(x) = \frac{1}{(2\pi)^{N/2}} \frac{1}{\sqrt{\det \mathcal{K}}} e^{-\frac{1}{2}(x-\mu)^T \mathcal{K}^{-1}(x-\mu)}, \quad x \in \mathbb{R}^N. \tag{3.2}$$

Definition 3.2 (Gaussian process) A stochastic process $\{X_\alpha\}_{\alpha \in A}$ is said to be a *Gaussian process* (GP) if every finite collection of random variables has a multivariate Gaussian distribution, i.e., for any finite subset $F \subseteq A$ there exist a vector $\mu_F \in \mathbb{R}^{|F|}$ and a positive semidefinite covariance matrix $\mathcal{K}_F \in \mathbb{R}^{|F| \times |F|}$ such that

$$\{X_\alpha\}_{\alpha \in F} \sim \mathcal{N}(\mu_F, \mathcal{K}_F). \tag{3.3}$$

In particular, if we define the mean function $\mu : A \to \mathbb{R}$ and a covariance function $\mathcal{K} : A \times A \to \mathbb{R}$ as

$$\mu(\alpha) = \mathbb{E}\left[X_\alpha\right], \tag{3.4}$$

$$\mathcal{K}(\alpha, \alpha') = \mathbb{E}\left[\big(X_\alpha - \mu(\alpha)\big)\big(X_{\alpha'} - \mu(\alpha')\big)\right], \tag{3.5}$$

then μ_F and K_F can be seen as the vector and the matrix produced by the evaluations of these functions:

$$\mu_F = \mu(F), \qquad \mathcal{K}_F = \mathcal{K}(F, F^T). \tag{3.6}$$

The mean function and the covariance function characterize the Gaussian process. Therefore, we will refer to the distribution of $\{X_\alpha\}_{\alpha \in A}$ by writing

$$\{X_\alpha\} \sim GP(\mu, \mathcal{K}). \tag{3.7}$$

As an example, the random output of a function $f : [0,1] \to \mathbb{R}$ can be distributed as a Gaussian process. In this case, we are setting $A = [0,1]$, and we are considering $\mu : [0,1] \to \mathbb{R}$ and $\mathcal{K} : [0,1] \times [0,1] \to \mathbb{R}$ so that

$$\{f(x)\}_{x\in[0,1]} \sim GP(\mu, \mathcal{K}). \tag{3.8}$$

Each sample of this distribution consists in a particular function f which is defined at each $x \in [0,1]$. If we repeated many times the sampling, we would notice that the shape of the function is statistically affected by $\mu(x)$ and $\mathcal{K}(x, x')$. In particular, if we fixed a point $x_1 \in [0,1]$ and we isolated the value $f(x_1)$ of each sample, we would notice that

$$\mathbb{E}[f(x_1)] = \mu(x_1). \tag{3.9}$$

A similar discussion can be done for the covariance of the values of the function and, more generally, for the distribution of any finite subset of values of the functions in the sample. The notion of Gaussian process, as in the classical setting, will be necessary to describe the output of the quantum circuit in the limit of infinitely many qubits.

In order to state the central limit theorem, we recall one of the equivalent definitions of the convergence in distribution (see e.g.[45, 46]).

Definition 3.3 (Convergence in distribution) Let X be a real valued random variable and let $X_1, X_2, X_3, \dots$ be a sequence of real-valued random variables. The sequence $X_1, X_2, X_3, \dots$ *converges in distribution* to X if

$$\lim_{k\to\infty} \mathbb{E}[f(X_k)] = \mathbb{E}[f(X)] \tag{3.10}$$

for any bounded and continuous function $f : \mathbb{R} \to \mathbb{R}$. We will use the notation

$$X_k \xrightarrow{d} X. \tag{3.11}$$

Lemma 3.1 (Equivalent definition of the convergence in distribution) *Let X be a real-valued random variable and let $X_1, X_2, X_3, \dots$ be a sequence of real-valued random variables. Then*

$$X_k \xrightarrow{d} X \quad \text{as} \quad k \to \infty \tag{3.12}$$

is equivalent to

$$\lim_{k\to\infty} \mathbb{P}\big(X_k \le x\big) = \mathbb{P}\big(X \le x\big) \quad \text{for any } x \in \mathbb{R} \text{ such that } x \mapsto \mathbb{P}\big(X \le x\big) \text{ is continuous.} \tag{3.13}$$

Proof See e.g. [45]. □

Theorem 3.2 (Central limit theorem) *Let $X_1, X_2, X_3, \dots$ be a sequence of i.i.d. random variables with*

$$\mathbb{E}[X_i] = 0 \qquad \mathrm{Var}[X_i] = \sigma^2 \tag{3.14}$$

and let $X \sim \mathcal{N}(0, \sigma^2)$. Then

$$\frac{1}{\sqrt{m}} \sum_{k=1}^{m} X_k \xrightarrow{d} X \quad \text{as} \quad m \to \infty. \tag{3.15}$$

Proof See [46]. □

The following result is used in the proof of the central limit theorem and we will need it to prove Theorem 3.7.

Theorem 3.3 (Lévy's continuity theorem) *Let $X_1, X_2, X_3, \dots$ be a sequence of random variable with characteristic functions*

$$\phi_k(t) = \mathbb{E}\left[e^{itX_k}\right] \qquad k \ge 1. \tag{3.16}$$

If there exists a random variable X whose characteristic function $\phi(t)$ is the pointwise limit of $\phi_k(t)$, i.e.

$$\lim_{k\to\infty} \phi_k(t) = \phi(t) \qquad \forall t \in \mathbb{R}, \tag{3.17}$$

then X_k converges in distribution to X:

$$X_k \xrightarrow{d} X. \tag{3.18}$$

Proof See e.g. [46–48]. □

Since, in our case, $f_k(\Theta, x)$ is not a sequence of i.i.d. random variables – the local observables are neither identically distributed nor independent – we need a strategy to generalize Theorem 3.2. This is the aim of the following paragraph.

3.1.2 Cumulants and dependency graphs

The strategy to prove Theorem 3.7 is based on the theory of cumulants of a random variable, which we briefly recall (see [49]):

Definition 3.4 (Cumulants) Given a real-valued random variable X, its *cumulants* $\kappa_j \equiv \kappa_j(X)$, $j \in \mathbb{N}$, are the defined as

$$\kappa_j(X) = (-i)^j \frac{d^j}{dt^j}\bigg|_{t=0} \log \mathbb{E}\left[e^{itX}\right], \tag{3.19}$$

provided that the derivative exists.

It is easy to prove that

$$\kappa_1(X) = \mathbb{E}[X], \qquad \kappa_2(X) = \mathrm{Var}[X], \tag{3.20}$$

when they exist. If the characteristic function is r-times differentiable, an equivalent characterization of the cumulants of order $j = 1, \dots, r$ is provided by the Taylor expansion:

$$\log \mathbb{E}\left[e^{itX}\right] = \sum_{j=1}^{r} \frac{\kappa_j}{j!}(it)^j + o(t^r) \qquad \text{as} \quad t \to 0. \tag{3.21}$$

Now we introduce the theory of dependency graphs following [50]. It will be the key tool to support and quantify the intuition that a circuit in which two local observables are "often" independent yields an output behaving as one would expect by the central limit theorem, even though the hypotheses are not satisfied.

Definition 3.5 Given a family of random variables $\{X_\alpha\}_{\alpha \in A}$, a graph $\mathscr{G}$ with vertex set V is called a *dependency graph* for the family if the following property holds: whenever V_1 and V_2 are disjoint subsets of V such that there are no edges in $\mathscr{G}$ with one end in V_1 and one in V_2, the subfamilies of random variables $\{X_\alpha\}_{\alpha \in V_1}$ and $\{X_\alpha\}_{\alpha \in V_2}$ are mutually independent. We stress that there is no assumption of independence within each of the sets $\{X_\alpha\}_{\alpha \in V_1}$ and $\{X_\alpha\}_{\alpha \in V_2}$.

The dependency graph of a family of random variables might not be unique: the complete graph is always a valid dependecy graph. However, as we will see in the next theorems, we look for dependency graph having as few edges as possible.

In Figure 12 we see an example of a dependency graph for a quantum circuit. The edges connect the index of the observables which are not mutually independent. We should prove that, *for our case*, this procedure to construct a dependency graph is compatible with Definition 3.5. Indeed, this statement is not generalizable to any family of random variables. An example of a triple of random variables which are independent in pairs but not globally independent can be found in [51, 52]. The case of the quantum circuit we are considering is not affected by this problem.

Lemma 3.4 *Let us consider a quantum circuit as in section 2 with m qubits. Let $\mathscr{G} = (V, A)$ be the graph with vertices $V = \{1, \dots, m\}$ and edges E defined by*

$$\begin{aligned} (k, k') \in E \qquad \Longleftrightarrow \qquad & f_k(\Theta, x) \text{ and } f_{k'}(\Theta, x) \\ & \text{are not independent at random initialization} \\ & \text{i.e., } k' \in \mathcal{P}_k. \end{aligned} \tag{3.22}$$

Then, $\mathscr{G}$ is a dependency graph for $\{f_k(\Theta, x)\}_{k \in V}$ according to Definition (3.5).

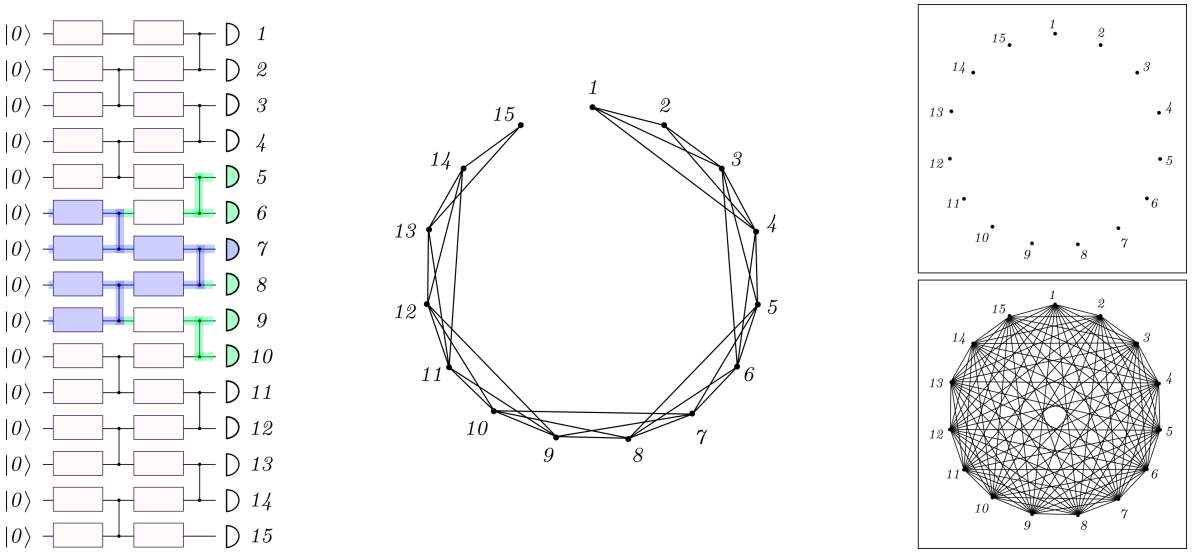


Fig. 12 A dependency graph for the circuit on the left compared with the trivial graph with no edges (independent observales) and the complete graph (all the variables are pairwise dependent). Informally speaking, the dependency graph of the circuit seems somehow closer to the trivial graph than the complete graph, since "far" observables are independent.

Remark 3.1 As already asserted in section 2, in all our discussion we assume that the parameters $\theta_1, \dots, \theta_m$ are *independent* random variables. We will not repeat every time this assumption, which will be almost always implicit.

Proof Recalling that

$$\begin{aligned}\mathcal{P}_k = \{k' \in \{1,\dots,m\} : f_{k'}(\Theta, x) \text{ is not independent of} \\ f_k(\Theta, x) \text{ at random initialization}\} \\ \subseteq \{k' \in \{1,\dots,m\} : \mathcal{N}_k \cap \mathcal{N}_{k'} \neq \emptyset\} \\ = \{k' \in \{1,\dots,m\} : \exists\, i \in \{1,\dots,Lm\} \text{ such that } k \in \mathcal{M}_i \text{ and } k' \in \mathcal{M}_i\}. \end{aligned} \tag{3.23}$$

Therefore, if $k' \notin \mathcal{P}_k$, then the observables

$$f_k(\Theta_{\mathcal{N}_k}, x) \qquad \text{and} \qquad f_{k'}(\Theta_{\mathcal{N}_{k'}}, x) \tag{3.24}$$

are functions of different parameters. Therefore, if V_1 and V_2 are subsets of V such that

$$\forall\, k \in V_1 \quad \nexists\, k' \in V_2 \quad \text{such that} \quad k' \in \mathcal{P}_k, \tag{3.25}$$

then the sets

$$\bigcup_{k \in V_1} \mathcal{N}_k \quad \text{and} \quad \bigcup_{k' \in V_2} \mathcal{N}_{k'} \tag{3.26}$$

are disjoint:

$$\left(\bigcup_{k \in V_1} \mathcal{N}_k\right) \cap \left(\bigcup_{k' \in V_2} \mathcal{N}_{k'}\right) = \bigcup_{\substack{k \in V_1 \\ k' \in V_2}} \mathcal{N}_k \cap \mathcal{N}_{k'} = \emptyset \tag{3.27}$$

since $\mathcal{N}_k \cap \mathcal{N}_{k'} = \emptyset$ for any $k \in V_1, k' \in V_2$. Noticing that the family $\{f_k(\Theta, x)\}_{k \in V_1}$ depends on $\bigcup_{k \in V_1} \mathcal{N}_k$ and the family $\{f_{k'}(\Theta, x)\}_{k' \in V_2}$ depends on $\bigcup_{k' \in V_2} \mathcal{N}_{k'}$, we conclude that they are independent, because they depend on different independent random variables. □

We define the *maximal degree* D of a graph $\mathscr{G} = (V, E)$ as the maximum number of edges incidenting on any fixed vertex

$$D = \max_{k \in V} \deg k = \max_{k \in V} |\{k' \in V : (k, k') \in E\}|. \tag{3.28}$$

Now that we are able to construct a dependency tree for our circuit, let us state the following technical theorem:

Theorem 3.5 ([50]) *For any integer $r \geq 1$, there exists a constant C_r with the following property. Let $\{X_\alpha\}_{\alpha \in V}$ be a family of random variables with dependency graph $\mathscr{G}$. We denote with $N = |V|$ the number of vertices of $\mathscr{G}$ and D the maximal degree of $\mathscr{G}$. Assume that the variables X_α are uniformly bounded by a constant A. Then, if*

$$X = \sum_{\alpha \in V} X_\alpha, \tag{3.29}$$

one has

$$|\kappa_r(X)| \leq C_r N (D+1)^{r-1} A^r \tag{3.30}$$

with

$$C_r = 2^{r-1} r^{r-2}. \tag{3.31}$$

The upper bound (3.30) was stated by Janson in [53]. Döring and Eichelsbacher established in [54] that the bound, according to the original proof, holds with $C_r = (2e)^r (r!)^3$. Féray, Méliot and Nikeghbali [50] found the new bound (3.31) by giving a new proof of (3.30).

3.2 Convergence to a Gaussian process

The following assumption fixes the expectation value of the model function at random initialization and the asymptotic behaviour of the normalization $N(m)$.

Assumption 3.6 *Where not otherwise specified, we will assume that the distribution of the independent random variables $\theta_1, \dots, \theta_{|\Theta|}$, the architecture of the quantum circuit and the normalization $N(m)$ chosen are such that*

$$\mathbb{E}[f_k(\Theta, x)] = 0 \qquad \forall\, x \in \mathcal{X} \tag{3.32}$$

and

$$\lim_{m \to \infty} \sup_{x, x' \in \mathcal{X}} \left| \mathbb{E}[f(\Theta, x) f(\Theta, x')] - \mathcal{K}(x, x') \right| = 0, \tag{3.33}$$

where $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is an arbitrary bivariate function from the feature space to the real numbers with strictly positive diagonal elements, i.e.

$$\mathcal{K}(x, x) > 0 \qquad \forall\, x \in \mathcal{X}. \tag{3.34}$$

Therefore $N(m)$ is chosen so that the limit $m \to \infty$ yields a finite nontrivial covariance function. This assumption does not fix a unique normalization, since a multiplicative constant or asymptotically negligible corrections to $N(m)$ produce other valid normalizations. Such ambiguity is irrelevant in the following discussion: after choosing the circuit we want to study, we will always assume to have fixed a possible normalization among the ones satisfying Assumption 3.6.

Remark 3.2 Assumption 3.6 implies that the operator norm of $\frac{\mathcal{O}}{N(m)}$ does in general scale with m. Indeed, let us consider the toy model where the expectation values of each $\mathcal{O}_k$ are independent. Then, to get a finite nonzero variance we need $N(m) = \sqrt{m}$, making the operator norm of $\frac{\mathcal{O}}{N(m)}$ scale as $\sqrt{m}$.

Remark 3.3 The architecture we introduced in section 2 satisfies (3.32) by Assumption 2.8.

Remark 3.4 The reader may be worried that the existence of the limit (3.33) may be a very restrictive hypothesis. However, we will now show that such limit always exists for any sequence of quantum neural networks satisfying the remaining hypotheses upon taking a suitable subsequence.

1. Due to the multiplicative degree of freedom in the definition of $N(m)$ for each m, we can suppose to fix a constant c (not depending on m) that uniformly bounds the matrix elements $|\mathbb{E}[f(\Theta, x) f(\Theta, x')]| \leq c$ for any $x, x' \in \mathcal{X}$.
2. Since we assumed that the set of input is finite (Assumption 2.1), by a standard compactness argument it is possible to identify a subsequence $(m_k)_{k \in \mathbb{N}}$ such that the limit

$$\mathcal{K}(x, x') \coloneqq \lim_{k \to \infty} \mathbb{E}[f^{(m_k)}(\Theta, x) f^{(m_k)}(\Theta, x')] \tag{3.35}$$

exists.

3. We can now study the properties of the circuits defined by this subsequence. To simplify the notation, in the rest of the paper we will not refer to any subsequence, but all our results can easily be generalized to such case.

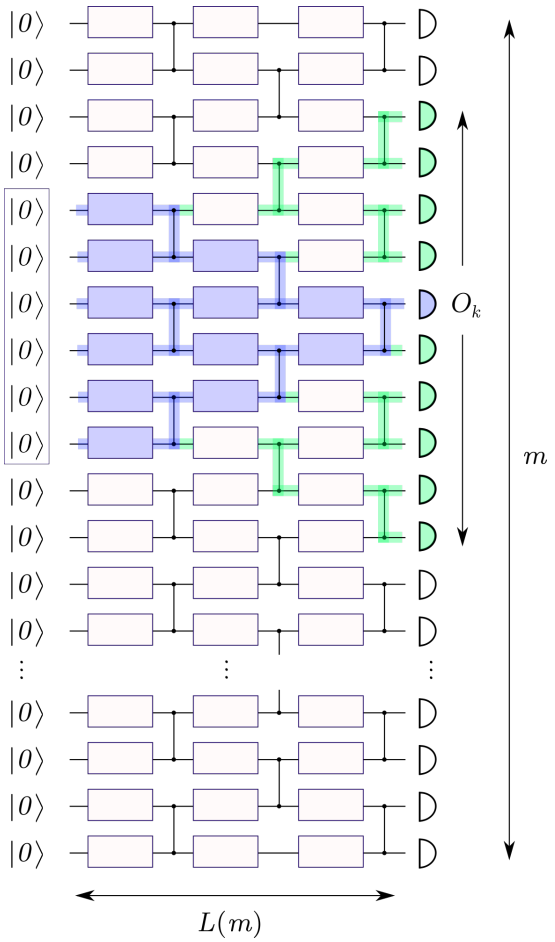


Fig. 13 For narrow enough light cones, the number of observables which are not independent of $f_k(\Theta, x)$ is small with respect to the total number of observables in the limit $m \to \infty$.

Now that for any $k \in \{1, \dots, m\}$ we can estimate the number of the observables which are not independent of $f_k(\Theta, x)$ at random initialization (Lemma 2.9) and that we know how to construct the dependency graph $\mathscr{G}$ of the quantum circuit (Lemma 3.4), we can give an upper bound to the maximal degree of $\mathscr{G}$ and apply then Theorem 3.5. Combining the result with a strategy analogous to the proof of the central limit theorem, we obtain the following Theorem, which is the main result of this section.

Theorem 3.7 (Gaussian process at initialization) *Given $\theta_1, \dots, \theta_{|\Theta|}$ independent random variables, suppose that Assumption 3.6 holds and that*

$$\lim_{m\to\infty} \frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3(m)} = 0. \tag{3.36}$$

Then, as $m \to \infty$, $x \mapsto f^{(m)}(\Theta, x)$ converges in distribution to a mean zero Gaussian process with covariance function $\mathcal{K}$ defined in Assumption 3.6:

$$f^{(m)}(\Theta, x) \xrightarrow{d} f^{(\infty)}(x) \sim GP(0, \mathcal{K}) \qquad m \to \infty. \tag{3.37}$$

3.3 Proof of Theorem 3.7

Before proving Theorem 3.7, we need to show that the hypothesis (3.36) implies a control on the growth of the light cones with respect to the normalization constant (see Corollary 3.9) due to an upper bound of the normalization constant (Lemma 3.8).

Lemma 3.8 *Assumption 3.6 implies the following bound: there exists $\bar{m} \in \mathbb{N}$ for any $m \geq \bar{m}$.*

$$N(m) \leq C\sqrt{m}\sqrt{|\mathcal{M}||\mathcal{N}|}. \tag{3.38}$$

Proof We recall that we assumed that $f_k(\Theta, x)$ are bounded observables:

$$|f_k(\Theta, x)| \leq 1 \qquad \forall k \in \{1, \dots, m\} \, \forall \, x \in \mathcal{X}, \, \forall \, \Theta \in \mathscr{P}. \tag{3.39}$$

By Assumption 3.6 there exists $\bar{m}_x \in \mathbb{N}$ such that, for any $m \geq \bar{m}_x$

$$0 < \frac{1}{2}\mathcal{K}(x, x) \leq \mathbb{E}[f^2(\Theta, x)]. \tag{3.40}$$

Hence, for $m \geq \bar{m}_x$, $\mathcal{K}(x, x)$ can be estimated as

$$\begin{aligned}
\frac{1}{2}\mathcal{K}(x, x) &\leq \frac{1}{N^2(m)} \sum_{k,k'=1}^{m} \mathbb{E}[f_k(\Theta, x) f_{k'}(\Theta, x)] \\
&= \frac{1}{N^2(m)} \sum_{k=1}^{m} \left(\sum_{k' \in \mathcal{P}_k} \mathbb{E}[f_k(\Theta, x) f_{k'}(\Theta, x)] + \sum_{k' \notin \mathcal{P}_k} \mathbb{E}[f_k(\Theta, x)]\mathbb{E}[f_{k'}(\Theta, x)] \right) \\
&= \frac{1}{N^2(m)} \sum_{k=1}^{m} \sum_{k' \in \mathcal{P}_k} \mathbb{E}[f_k(\Theta, x) f_{k'}(\Theta, x)] \\
&\leq \frac{1}{N^2(m)} \sum_{k=1}^{m} \sum_{k' \in \mathcal{P}_k} \mathbb{E}[|f_k(\Theta, x)||f_{k'}(\Theta, x)|] \\
&\leq \frac{1}{N^2(m)} \sum_{k=1}^{m} \sum_{k' \in \mathcal{P}_k} 1 \leq \frac{1}{N^2(m)} m \max_k |\mathcal{P}_k| \leq \frac{1}{N^2(m)} m|\mathcal{M}||\mathcal{N}|.
\end{aligned} \tag{3.41}$$

where the last inequality follows from Lemma 2.9. Let

$$C_x = \max \left\{ \frac{\sqrt{2}}{\sqrt{\mathcal{K}(x, x)}}, \frac{N(m)}{\sqrt{m}\sqrt{|\mathcal{M}||\mathcal{N}|}} \Big|_{m=1,\dots,\bar{m}_x - 1} \right\}. \tag{3.42}$$

With such a definition of C_x, the following inequality holds not only for $m \geq \bar{m}$, but more in general for any $m \geq 1$:

$$N(m) \leq C_x\sqrt{m}\sqrt{|\mathcal{M}||\mathcal{N}|} \qquad \forall \, x \in \mathcal{X}, \tag{3.43}$$

so

$$N(m) \leq C\sqrt{m}\sqrt{|\mathcal{M}||\mathcal{N}|} \qquad \forall \, m \geq 1, \tag{3.44}$$

where

$$C = \inf_{x \in \mathcal{X}} C_x. \tag{3.45}$$

□

Corollary 3.9 *If*

$$\lim_{m \to \infty} \frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3(m)} = 0, \tag{3.46}$$

then

$$\lim_{m \to \infty} \frac{|\mathcal{M}||\mathcal{N}|}{N(m)} = 0. \tag{3.47}$$

Proof First, by Lemma 3.8,

$$\frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3(m)} \geq \frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{C^3 m^{3/2}|\mathcal{M}|^{3/2}|\mathcal{N}|^{3/2}} = \frac{1}{C^3}\frac{\sqrt{|\mathcal{M}||\mathcal{N}|}}{\sqrt{m}} \tag{3.48}$$

which implies

$$\lim_{m\to\infty} \frac{|\mathcal{M}||\mathcal{N}|}{m} = 0. \tag{3.49}$$

Besides, using again Lemma 3.8

$$1 \leq C\frac{\sqrt{|\mathcal{M}||\mathcal{N}|}\sqrt{m}}{N(m)}, \tag{3.50}$$

therefore

$$\begin{aligned}\frac{|\mathcal{M}||\mathcal{N}|}{N(m)} &= \left(\frac{|\mathcal{M}|^2|\mathcal{N}|^2}{N^2(m)}\right)^{1/2} \leq \left(C\frac{\sqrt{|\mathcal{M}||\mathcal{N}|}\sqrt{m}}{N(m)}\frac{|\mathcal{M}|^2|\mathcal{N}|^2}{N^2(m)}\right)^{1/2} \\ &= \sqrt{C}\left(\frac{|\mathcal{M}||\mathcal{N}|}{m}\right)^{1/4}\left(\frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3(m)}\right)^{1/2},\end{aligned} \tag{3.51}$$

whence

$$\lim_{m\to\infty}\frac{|\mathcal{M}||\mathcal{N}|}{N(m)} \leq \sqrt{C}\lim_{m\to\infty}\left(\frac{|\mathcal{M}||\mathcal{N}|}{m}\right)^{1/4}\lim_{m\to\infty}\left(\frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3(m)}\right)^{1/2} = 0. \tag{3.52}$$

□

We now have all the ingredients to prove Theorem 3.7.
We fix a finite collection $x_A = (x^{(\alpha)}, \alpha \in A)$. Let $\xi \in \mathbb{R}^{|A|}$ and $\|\xi\|_1 = \sum_\alpha |\xi_\alpha|$. The characteristic function of the finite collection of random variables $f^{(m)}(\Theta, x_A)$ is

$$\begin{aligned}\phi_m(\xi) &= \mathbb{E}\left[\exp\left\{if^{(m)}(\Theta, x_A)\cdot\xi_A\right\}\right] \\ &= \mathbb{E}\left[\exp\left\{i\frac{1}{N(m)}\sum_{k=1}^m f_k(\Theta, x_A)\cdot\xi_A\right\}\right],\end{aligned} \tag{3.53}$$

where

$$f_k(\Theta, x_A)\cdot\xi_A \equiv \sum_{\alpha\in A} f_k(\Theta, x^{(\alpha)})\cdot\xi_\alpha. \tag{3.54}$$

We rename

$$Y^{(m)} \equiv \frac{1}{N(m)}\sum_{k=1}^m f_k(\Theta, x_A)\cdot\xi_A \tag{3.55}$$

and we introduce a fictitious variable t such that $\phi_m(\xi) = \phi_m(\xi, t=1)$:

$$\begin{aligned}\phi_m(\xi, t) &= \mathbb{E}\left[\exp(itf^{(m)}(\Theta, x_A)\cdot\xi_A)\right] \\ &= \mathbb{E}\left[\exp(itY^{(m)})\right].\end{aligned} \tag{3.56}$$

As we saw in paragraph 3.1.2, $\log\phi_m(\xi, t)$ can be expanded in terms of the cumulants of $Y^{(m)}$:

$$\kappa_r^{(m)} \equiv (-i)^n\frac{d^r}{dt^r}\log\phi_m(\xi, t) \qquad \to \qquad \log\phi_m(\xi, t) = \sum_{r=1}^\infty \frac{\kappa_r^{(m)}}{r!}(it)^r. \tag{3.57}$$

As we proved in Lemma 3.4, a valid dependency graph $\mathscr{G} = (V, E)$ of the quantum circuit is

$$V = \{1, 2, \ldots, m\} \qquad E = \{(k, k') \in V\times V : k' \in \mathcal{P}_k\}, \tag{3.58}$$

where

$$\mathcal{P}_k = \{k' \in \{1,\dots,m\} : f_{k'}(\Theta, x) \text{ is not independent of } f_k(\Theta, x) \text{ at random initialization}\}. \tag{3.59}$$

Calling

$$D = \max_{k\in V} \deg k, \tag{3.60}$$

by means of Lemma 2.9 we can bound

$$D = \max_{k\in V} \left|\{(k,k') \in V \times V \,:\, k' \in \mathcal{P}_k\}\right| = \max_{k\in I} |\mathcal{P}_k| \le |\mathcal{M}||\mathcal{N}|. \tag{3.61}$$

The random variable $Y^{(m)}$ is a sum of uniformly bounded random variables by Assumption 2.2:

$$\left|\frac{1}{N(m)} f_k(\Theta, x_A)\cdot \xi_A\right| \le \frac{\|\xi\|_1}{N(m)} \tag{3.62}$$

Therefore, Theorem 3.5 can be applied:

$$|\kappa_r^{(m)}| \le 2^{r-1} r^{r-2} m (D+1)^{r-1} \left(\frac{\|\xi\|_1}{N(m)}\right)^r \le \frac{m}{D}\left(\frac{4\|\xi\|_1 D}{N(m)}\right)^r r^r. \tag{3.63}$$

Using the following inequality for the factorial [55]

$$e\frac{r^r}{e^r} \le r! \qquad \to \qquad r^r \le e^r r!, \tag{3.64}$$

we further estimate

$$|\kappa_r^{(m)}| \le \frac{m}{D}\left(\frac{4e\|\xi\|_1 D}{N(m)}\right)^r r!. \tag{3.65}$$

In this way, we can control the tail $T(t,\xi)$ of the Taylor expansion of $\log \phi_m(\xi, t)$:

$$T(t,\xi) := \sum_{r=3}^{\infty} \frac{\kappa_r^{(m)}}{r!}(it)^r, \tag{3.66}$$

$$\begin{aligned} |T(t,\xi)| = \left|\sum_{r=3}^{\infty} \frac{\kappa_r^{(m)}}{r!}(it)^r\right| &\le \sum_{r=3}^{\infty} \frac{|\kappa_r^{(m)}|}{r!} t^r \\ &\le \sum_{r=3}^{\infty} \frac{m}{D}\left(\frac{4e\|\xi\|_1 tD}{N(m)}\right)^r \\ &= (4e\|\xi\|_1 t)^3 \frac{mD^2}{N^3(m)} \sum_{r=0}^{\infty}\left(\frac{4e\|\xi\|_1 tD}{N(m)}\right)^r \\ &\le (4e\|\xi\|_1 t)^3 \frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3(m)} \sum_{r=0}^{\infty}\left((4e\|\xi\|_1 t)\frac{|\mathcal{M}||\mathcal{N}|}{N(m)}\right)^r. \end{aligned} \tag{3.67}$$

The following bounds hold for fixed t and ξ. If, by hypothesis,

$$\lim_{m\to\infty} \frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3(m)} = 0, \tag{3.68}$$

then, by Corollary 3.9, we must also have

$$\lim_{m\to\infty} \frac{|\mathcal{M}||\mathcal{N}|}{N(m)} = 0, \tag{3.69}$$

which implies, for some m_0 (which depends on t and ξ),

$$(4e\|\xi\|_1 t)\frac{|\mathcal{M}||\mathcal{N}|}{N(m)} \leq \frac{1}{2} \qquad \forall\, m \geq m_0. \tag{3.70}$$

Therefore

$$|T(t,\xi)| \leq \frac{(4e\|\xi\|_1 t)^3}{2}\frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3(m)} \qquad \forall\, m \geq m_0 \qquad \to \qquad \lim_{m\to\infty} T(t,\xi) = 0. \tag{3.71}$$

So we can expand

$$\log \phi_m(\xi) = \log \phi_m(\xi, 1) = \kappa_1^{(m)}(\xi) - \frac{1}{2}\kappa_2^{(m)}(\xi) + T(\xi, 1). \tag{3.72}$$

Recalling that A is finite and using the summation convention over α and β, we can compute

$$\kappa_1^{(m)} = \mathbb{E}[Y^{(m)}] = 0 \text{ because of the (3.32)}, \tag{3.73}$$

$$\kappa_2^{(m)} = \mathbb{E}[(Y^{(m)})^2] - \mathbb{E}[Y^{(m)}]^2 = \frac{1}{N^2(m)} \sum_{k,k'=1}^{m} \xi_\alpha\, \mathbb{E}[f_k(\Theta, x^{(\alpha)}) f_{k'}(\Theta, x^{(\beta)})]\, \xi_\beta, \tag{3.74}$$

$$\to \kappa_2^{(m)} \to \xi_\alpha\, \mathcal{K}(x^{(\alpha)}, x^{(\beta)})\, \xi_\beta \quad \text{ as } m \to \infty \text{ because of the (3.33)}. \tag{3.75}$$

Since the logarithm is a continuous function, we have pointwise convergence:

$$\lim_{m\to\infty} \phi_m(\xi) = \exp\left\{-\frac{1}{2}\xi_\alpha \mathcal{K}(x^{(\alpha)}, x^{(\beta)})\xi_\beta\right\}. \tag{3.76}$$

This is the the characteristic function of a multivariate random variable (whose distribution is jointly Gaussian), so we can conclude by Lévy's theorem 3.3.

Remark 3.5 (A counterexample for large light cones) The hypotheses under which Theorem 3.7 holds are very general, since the proof is based only on the dependency relations between the observables. Many studied examples of circuits respecting the most general structure of a variational circuit could be considered, even with arbitrary (and artificial) parameter dependence and distribution. It is interesting to notice that, if we relax the hypothesis on the growth of the light cones, we can exhibit a circuit that violates Theorem 3.7. An explicit construction is done in Appendix A.

4 Trained quantum neural networks are Gaussian processes

4.1 Introduction

4.1.1 Training quantum neural networks

As in the classical setting, the aim of the training is the minimization of a cost function $\mathcal{L}(\Theta, \mathcal{D})$ (from now on, we will omit the dependence on $\mathcal{D}$) based on the discrepancy between the model function $f(\Theta, x^{(i)})$ and the training labels $y^{(i)}$. This can be done in continuous time by the gradient flow method. The vector of the parameters evolves according to the differential equation

$$\dot{\Theta}_t = -\eta \nabla_\Theta \mathcal{L}(\Theta_t), \tag{4.1}$$

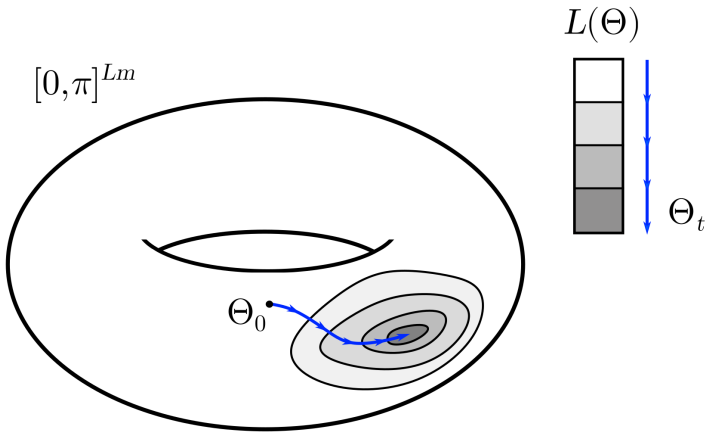


Fig. 14 Gradient flow in the parameter space

where $\eta > 0$ is a constant which resizes the time. This equation ensures that the cost is decreasing during the evolution (see Figure 14)

$$\frac{d}{dt}\mathcal{L}(\Theta_t) = \dot{\Theta}_t \cdot \nabla_\Theta \mathcal{L}(\Theta_t) = -\eta \|\nabla_\Theta \mathcal{L}(\Theta_t)\|_2^2 \leq 0. \tag{4.2}$$

The problem of the minimization of the cost is, in general, non-convex, so the convergence to a global minimum is not ensured. We are going to discuss this issue in subsubsection 4.1.2.

Since the computation of the model function in a concrete setting happens in discrete time, the minimizing procedure will be by described by the gradient descent method, which consists in the update of the parameters according to the rule

$$\Theta_{t+1} - \Theta_t = -\eta \nabla_\Theta \mathcal{L}(\Theta_t). \tag{4.3}$$

Besides, we should notice that the right-hand side of (4.3) is exactly known only performing an infinite amount of measurements, since the model function is defined as an expectation value (and, at the moment, we do not have a procedure to compute the derivatives of the model function). In a real setting, we will be able to perform only a finite amount of measurements of the output of the circuit. Therefore, we will need to substitute the RHS with an estimator of the gradient.
In this section, in order to understand the behaviour of the circuit during the training procedure and to focus on the main properties of the limit $m \to \infty$, we will start considering the gradient flow equation without noise. Later, in the next section, we will understand how to deal with the discrete time and the noise, and we will prove that the limit $m \to \infty$ allows us to solve a non-convex optimization problem.

Remark 4.1 We stress that the parametric unitary operator $U(\Theta, x)$ encodes both the input x and the trainable parameters Θ. Therefore, the training does not generate a single state of m qubits, but rather a parametric family of states $\{U(\Theta, x)|0^m\rangle\}_{x\in\mathcal{X}}$ that depend on the input x. For this reason, training the quantum neural networks considered here is not equivalent to finding a low-energy state of a given Hamiltonian, and the two tasks may have very different complexities. Therefore, results on the classical easiness of finding low-energy states such as [56] may not apply in the scenario considered in this paper.

4.1.2 Trainability: the loss landscape and barren plateaus

Before entering into the details of the equations of our model, let us discuss briefly the possible landscapes that may appear in a generic optimization task and the consequences on the gradient-flow training. We will also mention the main problems discussed in the current literature.

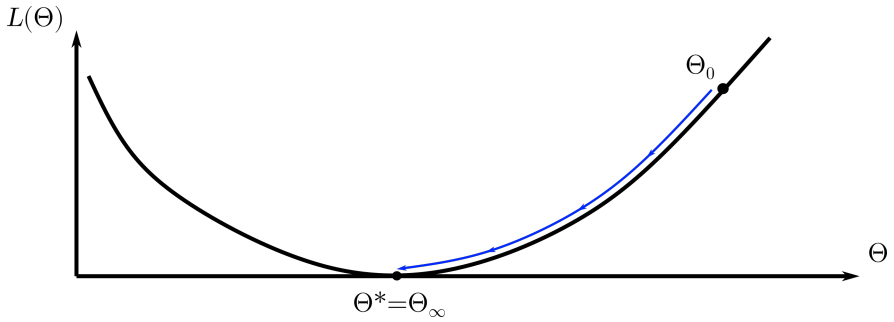


Fig. 15 Convex $\mathcal{L}(\Theta)$.

When $\mathcal{L}$ is a convex[8] function, then the convergence to the minimum is guaranteed, whatever the result of the initialization is (see Figure 15).

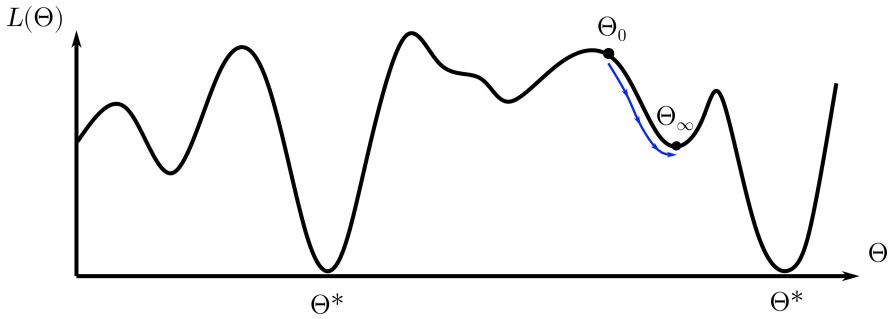


Fig. 16 Non-convex $\mathcal{L}(\Theta)$.

On the contrary, a generic non-convex cost function may be characterized by maxima, saddle points, local minima which are stationary points for the gradient flow evolution. In particular, after a random initialization in which the initial set of parameters is sampled far from a global minimum, the gradient flow evolution could lead to a local minimum, where the naive algorithm stops: the cost function cannot be further minimized by a continuous evolution of the parameters. Therefore, the result is not the one expected (see Figure 16).

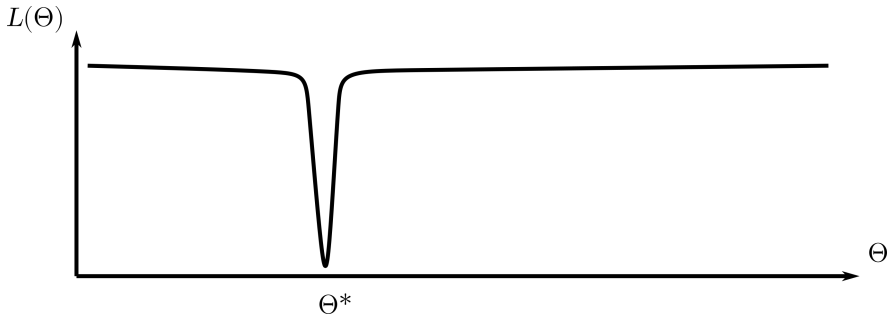


Fig. 17 Barren plateaus scenario.

Different problems may arise in the training of a quantum circuit. The main problem is constituted by barren plateaus (see e.g. [14–16, 58]), depicted in Figure 17: in some classes of circuits, typically when they are very deep, after a random initialization of the parameters the optimization landascape is characterized by gradients with a norm that decays exponentially with the number of qubits. This means that the convergence of the training is slow and that the estimate of the expectation value of the

[8]There are various shades of convexity, which can provide different results concerning the rate of convergence of the minimization procedure. For a survey on the main definitions, see [57].

gradient, in order to be precise enough, requires an exponentially large amount of measurements. This would hinder any potential speedup of the quantum algorithm.

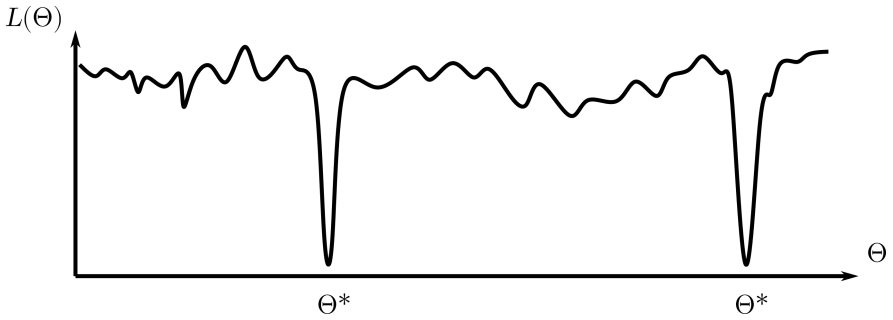


Fig. 18 Swampland of local minima.

There are other possible obstacles: in some class of shallow circuits which does not exhibit barren plateaus, the loss landscape can be swamped with many local minima which are far from the global minimum [13] (see Figure 18). If the local minima are rare with respect to global minima, the initialization procedure and the consequent training can be iterated until the global minimum is reached. But if the landscape is dense of local minima, we may have a too large number of iterations necessary to find a starting point for which the training converges to optimal parameters.

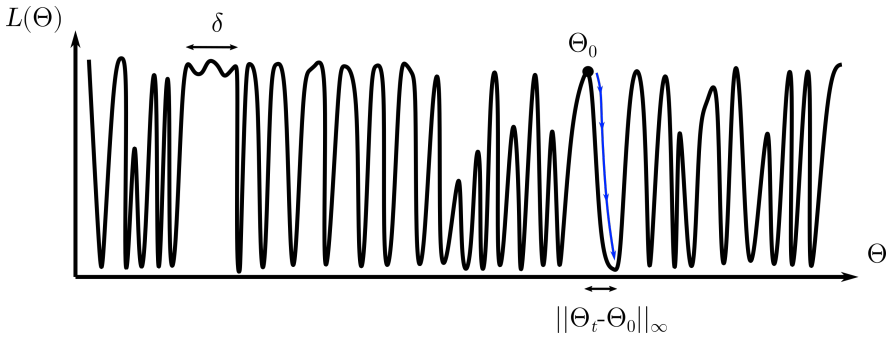


Fig. 19 Our scenario: lazy training.

We are going to prove that the limit $m \to \infty$ yields a very peculiar landscape (see Figure 19), which has many features in common with the one of wide classical neural networks. First, the minimum of the empirical cost approaches to zero as the number of qubit increases: the model becomes able to fit the dataset. Second, fixed any $\delta > 0$, with probability at least $1-\delta$ over random initialization, the parameters evolve succesfully from the starting point Θ_0 to a minimizer Θ^* which is increasingly close to Θ_0 as $m \to \infty$ (lazy training). The closeness of Θ_0 and Θ^* allows to compare the model with its first-order Taylor expansion around Θ_0. This comparison is motivated by the fact that an analytic solution for the linearized model is computable and its probability distribution during training is known.

In order to achieve these results – and to generalize them to the noisy setting in section 5 –, we follow a strategy which combines some ideas from [22], [38] and [59].

1. In subsection 4.2, we translate the gradient flow equation (4.19) into a differential equation for the model function (4.22) and we define the empirical neural tangent kernel (Definition 4.1), which is a bivariate symmetric function of the input space appearing in (4.22) and describing the evolution of the randomly initialized network during the training.
2. In subsection 4.3, we define the analytic neural tangent kernel as the expectation value of the empirical neural tangent kernel (Definition 4.2) and we study its properties.
3. In subsection 4.4, we introduce the linearized version – which is the first order Taylor expansion – of the model (4.34) and we compute its evolution equations under gradient flow.

4. In the first part of subsection 4.5, we show that the empirical neural tangent kernel with the parameters initialized randomly concentrates on its expectation value (Theorem 4.7), using a strategy which generalizes the proof of [38]. Differently from [38], our bounds are quantitative and they hold not only in the case with a fixed number of layers and with a geometrically local architecture, but for any dependence of the number of layers with respect to the number of qubits and for any architecture. As a consequence of our concentration inequality, we show that the linear model is distributed as a Gaussian process during all the evolution in the limit of infinitely many qubits (Corollary 4.9).
5. In the second part of subsection 4.5, we generalize the classical proof of [22] to the quantum setting in order to show that, as the number of qubits increases, the trained model converges to the examples – i.e., the cost function converges to zero – and the parameters are increasingly close to their value at initialization (Theorem 4.10).
6. In the last part of subsection 4.5, with some computations involving the size of the light cones, we show that the discrepancy between the model and its linearized version is small if the parameters are close to their value at initialization (Theorem 4.23). As a consequence, we show that the model converges to its linearized version in the limit of infinitely many qubits (Theorem 4.12). This result is a strong generalization of Theorem 4.2 of [38]: not only we consider the more general setting of a non-fixed number of layers and of generic architectures, but we also extend the convergence of the model to the linearized version from the dataset to all the input space. Furthermore, in [38] the convergence to the linearized model at any time can be achieved only by performing first the limit of infinitely many qubits, while for a circuit with finite size their bound diverges as t^2; clearly it is a fictitious divergence, since the training is convergent. Our bounds are valid for any time also in the (physical) finite size case.
7. As a consequence of the previous results, we conclude that the model function converges to a Gaussian process in the limit of infinitely many qubits (Theorem 4.15). This is the main result of this section.

Most of these strategies are valid also in the noisy setting with a discrete time evolution, provided that the variance of the estimator of the gradient is sufficiently small. We will prove in section 5 that there is a polynomial[9] upper bound on the variance which ensures an exponentially fast convergence of the training as in the simpler setting of this section.

4.1.3 Lazy training in the quantum setting

Before presenting our results, let us briefly review the results of a recent work about lazy training for quantum neural networks. In particular, we will focus on the statements and the bounds that we need to improve in order to achieve our convergence results.

E. Abedi, S. Beigi and L. Taghavi study in [38] a generalization to the quantum setting of the results of [60] for the classical case. In particular, they consider a geometrically local circuit[10] having a fixed number of layers. They prove the following results, which hold in the limit of infinitely many qubits:

1. the neural tangent kernel (see (4.8)) concentrates on its expectation value at random initialization (Theorem 4.1);
2. the evolution of the parameters is small during the training, so the linearized model behaves similarly to the original model (Theorem 4.2).

Let us quote the precise statement of the theorems proved by [38]. Differently from the notation we used so far, the parameters are labelled $\theta_1, \dots, \theta_p$ in [38].

Theorem 4.1 ([38]) *Let $f(\Theta, x)$ be a model function associated to a geometrically local parameterized quantum circuit on m qubits with a fixed number of layers*

$$f(\Theta, x) = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} f_k(\Theta, x), \qquad f_k(\Theta, x) = \langle 0^m | U^\dagger(\Theta, x) \mathcal{O}_k U(\Theta, x) | 0^m \rangle, \tag{4.4}$$

such that

$$\left\| \frac{\partial}{\partial \theta_j} U(\Theta, x) \right\| \leq c, \qquad \left\| \frac{\partial^2}{\partial \theta_i \partial \theta_j} U(\Theta, x) \right\| \leq c \qquad \forall i, j \in \{1, \dots, |\Theta|\}. \tag{4.5}$$

[9] with respect to the number of qubits

[10] More precisely, [38] requires that *the qubits are arranged on nodes of a bounded-degree graph*, that *the two-qubit gates can be applied only on pair of qubits connected by an edge* and that *the observable is a sum of terms, each of which acts only on a constant number of neighboring qubits.*

Suppose that the observable $\mathcal{O}$ is also geometrically local given by

$$\mathcal{O} = \sum_{k=1}^{m} \mathcal{O}_k, \tag{4.6}$$

where $\mathcal{O}_k$ acts on the k-th qubit and possibly on a constant number qubits in its neighborhood, and satisfies $\|\mathcal{O}_k\|_{\mathcal{L}} \leq 1$. In this case the model function is given by (4.4). Suppose that $\theta_1, \ldots, \theta_p$ are chosen independently at random. Then, for any $x, x' \in \mathbb{R}^d$ we have

$$\mathbb{P}\left(|\hat{K}_\Theta(x,x') - \mathbb{E}[\hat{K}_\Theta(x,x')| \geq \epsilon\right) \leq \exp\left(-\Omega\left(\frac{m^2\epsilon^2}{pc^4}\right)\right), \tag{4.7}$$

where

$$\hat{K}_\Theta(x,x') := \nabla_\Theta f(\Theta, x) \cdot \nabla_\Theta f(\Theta, x'), \tag{4.8}$$

Remark 4.2 We will introduce a normalization $N_K(m)$, depending on the number of qubits, in the definition of the (empirical) neural tangent kernel $\hat{K}_\Theta(x,x')$ (see Definition 4.21).

Theorem 4.2 ([38]) *Let $f(\Theta, x)$ be a model function associated with a parameterized quantum circuit satisfying the assumptions of Theorem 4.1. Suppose that a dataset $\mathcal{D}$*

$$\mathcal{D} = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\} \quad \text{with} \quad x^{(i)} \in \mathbb{R}^d \text{ and } y^{(i)} \in \mathbb{R} \tag{4.9}$$

is given. Assume that at initialization we choose $\Theta_0 = (\theta_1(0), \ldots, \theta_p(0))$ independently at random, and apply the gradient flow to update the parameters in time by

$$\dot{\Theta}_t = -\nabla_\Theta \mathcal{L}(\Theta_t), \tag{4.10}$$

where $\mathcal{L}(\Theta)$ is the mean squared error. Then, the followings hold.

1. *For any $1 \leq j \leq p$ we have*

$$|\partial_t \theta_j(t)| = O\left(\sqrt{\frac{\mathcal{L}(\Theta_0)}{m}}\right). \tag{4.11}$$

2. *For any x, x' we have*

$$|\partial_t \hat{K}_{\Theta_t}(x,x')| = O\left(\sqrt{\frac{\mathcal{L}(\Theta_0)}{m}}\right). \tag{4.12}$$

3. *Let $f^{\text{lin}}(\Theta, x)$ be the function associated to the linearized model, i.e.*

$$f^{\text{lin}}(\Theta, x) = f(\Theta_0, x) + \nabla_\Theta f(\Theta_0, x) \cdot (\Theta - \Theta_0). \tag{4.13}$$

Suppose we train the linearized model with its associated loss function

$$\begin{cases} \dfrac{d}{dt}\Theta_t^{\text{lin}} = -\nabla_\Theta \mathcal{L}^{\text{lin}}(\Theta_t^{\text{lin}}) \\ \Theta_0^{\text{lin}} = \Theta_0, \end{cases} \qquad \mathcal{L}^{\text{lin}}(\Theta_t) = \frac{1}{n}\sum_{i=1}^{n}\left(f^{\text{lin}}(\Theta, x^{(i)}) - y^{(i)}\right)^2. \tag{4.14}$$

Then, for all t we have

$$\left(\frac{1}{n}\sum_{i=1}^{n}\left(f(\Theta_t, x^{(i)}) - f^{\text{lin}}(\Theta_t^{\text{lin}}, x^{(i)})\right)^2\right)^{1/2} = O\left(\frac{\mathcal{L}(\Theta_0)t^2}{\sqrt{m}}\right). \tag{4.15}$$

and

$$|\mathcal{L}(\Theta_t) - \mathcal{L}^{\text{lin}}(\Theta_t^{\text{lin}})| = O\left(\frac{\mathcal{L}(\Theta_0)^{3/2}t^2}{\sqrt{m}}\right). \tag{4.16}$$

We improve both these results:

1. We will consider a more general setting in which the number of layers is not fixed and can grow with the number of qubits. This is a necessary condition to achieve a quantum advantage, as we discussed in section 2. Furthermore, our discussion will not be restricted to the class of geometrically local circuits: the only locality assumption we will state will concern the observable measured at the end of the circuit. Therefore, we will generalize (4.7) to our larger class of circuits and we will provide a quantitative bound (Theorem 4.7): the constant c appearing in (4.48) can be explicitly computed following the proof we give.
2. As remarked in [38], the bounds of Theorem 4.2 are effective when the cost function at initialization does not depend on the number of qubits. More precisely, it is enough that $\mathcal{L}(\Theta_0) = o(m^{1/3})$ in order not to trivialize (4.12), (4.12), (4.15) and (4.16) as $m \to \infty$. [38] claims to expect this property when the model at initialization approaches a Gaussian process, but no proof concerning the probability distribution of the function is provided. In section 3 we have proved that, in the limit of infinitely many qubits, the function generated by a quantum neural network with randomly initialized parameters converges to a Gaussian process when the parameters on which each qubit depends influence only a small number of other qubits (Theorem 3.7). Therefore, the cost function at initialization is bounded, with high probability, by a constant (Corollary 4.16).
3. We will prove that the whole training takes place in the lazy regime: the displacement of each parameter from its initial value is bounded by quantity which is convergent as $t \to \infty$ for m fixed, and which is arbitrarily small as $m \to \infty$ (Theorem 4.10), i.e.

$$\sup_t \|\Theta_t - \Theta_0\|_\infty \leq C(m), \qquad \text{with} \qquad \lim_{m\to\infty} C(m) = 0. \tag{4.17}$$

   A bound on the derivative of the form (4.11) does not imply the stronger property (4.17). The same improvement will be done for (4.12) in Corollary 4.11: we will prove that the neural tangent kernel is frozen on its value at initialization.
4. Furthermore, the bounds (4.15) and (4.16) become trivial as $t \to \infty$ with m fixed. We will prove a quantitative bound for the distance between the original model and its linear approximation *over the entire input space* with no divergences for $t \to \infty$ and m fixed (Theorem 4.12).

4.2 Gradient flow and neural tangent kernel

We call Θ_t the evolution of Θ_0 under gradient flow. We will sometimes use the compact notation

$$F(t) = \begin{pmatrix} f(\Theta_t, x^{(1)}) \\ f(\Theta_t, x^{(2)}) \\ \vdots \\ f(\Theta_t, x^{(n)}) \end{pmatrix}, \quad \text{where} \quad (x^{(i)}, \cdot) \in \mathcal{D} \quad n = |\mathcal{D}| \tag{4.18}$$

or similar vectorized forms of the model function. Using the gradient flow equation

$$\dot{\Theta}_t = -\eta \nabla_\Theta \mathcal{L}(\Theta_t), \tag{4.19}$$

and the chain rule, the equations for the evolution of the parameters and the model function can be rewritten as follows

$$\begin{cases} \dot{\Theta}_t &= -\eta \nabla_\Theta f(\Theta_t, X^T) \nabla_{f(\Theta_t, X)} \mathcal{L}(\Theta_t) \\ \dfrac{d}{dt} f(\Theta_t, x) &= \big(\nabla_\Theta f(\Theta_t, x)\big)^T \dot{\Theta}_t = -\eta \big(\nabla_\Theta f(\Theta_t, x)\big)^T \nabla_\Theta f(\Theta_t, X^T) \nabla_{f(\Theta_t, X)} \mathcal{L}(\Theta_t) \end{cases}. \tag{4.20}$$

Definition 4.1 (Empirical NTK) We define the *empirical neural tangent kernel* (NTK) as

$$\hat{K}_\Theta(x, x') = \frac{1}{N_K(m)} \big(\nabla_\Theta f(\Theta, x)\big)^T \nabla_\Theta f(\Theta, x'), \tag{4.21}$$

where $N_K(m)$ is a normalization factor, which will be specified by Assumption 4.3.

The set of equations (4.20) becomes

$$\begin{cases} \dot{\Theta}_t &= -\eta \nabla_\Theta f(\Theta_t, X^T) \nabla_{f(\Theta_t, X)} \mathcal{L}(\Theta_t) \\ \frac{d}{dt} f(\Theta_t, x) &= -\eta \, N_K(m) \, \hat{K}_{\Theta_t}(x, X^T) \nabla_{f(\Theta_t, X)} \mathcal{L}(\Theta_t) \end{cases} . \tag{4.22}$$

4.3 Assumptions and properties of the NTK

In the previous subsection we introduced the empirical NTK. Now, we introduce its expectation value over random initialization of the parameters and we discuss some properties.

Definition 4.2 (Analytic NTK) The *analytic NTK* K is defined as the expectation value of the empirical NTK $\hat{K}_\Theta$ over random initialization of the parameters:

$$K(x, x') \equiv \mathbb{E}\left[\hat{K}_\Theta(x, x')\right]. \tag{4.23}$$

Assumption 4.3 *We assume that we can choose a normalization $N_K(m)$ such that there exists a limit function $\bar{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$*

$$\lim_{m \to \infty} \sup_{x, x' \in \mathcal{X}} |K(x, x') - \bar{K}(x, x')| = 0 \tag{4.24}$$

not identically zero. We also assume that the minimum eigenvalue of the finite[11] *matrix $\bar{K} \equiv \bar{K}(X, X^T)$ is strictly positive: $\lambda^K_{\min} > 0$.*

Remark 4.3 We stress that, provided that Assumptions 3.6 and 4.3 are satisfied, the architecture of the quantum neural networks can be completely arbitrary. In particular, we do not need to assume that the qubits are arranged on a regular lattice with 2-qubit gates acting only between neighboring qubits.

Remark 4.4 Proceeding as in Remark 3.4, the limit (4.24) always exists upon taking a suitable subsequence.

Since $\bar{K}$ is a finite matrix, its spectrum is bounded. In particular, we will call $\lambda^K_{\max}$ its maximum eigenvalue.

As in the case of $N(m)$, we are left with some degrees of freedom in the choice of $N_K(m)$ under Assumption 4.3: a global factor or negligible corrections do not change that requirement. As we will see, a multiplicative factor $\alpha > 0$ in

$$K \quad \to \quad \alpha K \tag{4.25}$$

would rescale

$$\lambda^K_{\min} \quad \to \quad \alpha \lambda^K_{\min}. \tag{4.26}$$

Since $\lambda^K_{\min}$ will regulate the rate of convergence of the training (see Theorem 4.10), in the face of this rescaling one may question the well-definedness of this quantity. As we will see, an arbitrary parameter η_0 will always multiply $\lambda^K_{\min}$ in the expression of the rate of convergence: therefore, any rescaling constant can be absorbed in η_0, and this simply represents a reparametrization of the continuous time in the gradient flow equation.

Assumption 4.4 *From now on, we will assume that the parameters $\{\theta_i\}_{i=1,\dots,Lm}$ are distributed as independent uniform random variables in $[0, \pi]$, i.e., the initialization of the circuit is uniform in the parameter space $\mathscr{P} = [0, \pi]^{Lm}$.*

Under Assumption 4.4, we can more explicitly write

$$K(x, x') = \frac{1}{N_K(m)} \int_{[0,\pi]^{Lm}} \frac{d^{Lm}\theta}{(\pi)^{Lm}} \sum_i \partial_{\theta_i} f(\Theta, x) \partial_{\theta_j} f(\Theta, x). \tag{4.27}$$

[11] The number of examples in $\mathcal{D}$ is assumed to be finite.

Remark 4.5 As we will prove later (see Theorem 4.15), the analytic NTK, together with the training set and the covariance at initialization, fully characterizes the probability distribution of the function generated by the quantum neural network in the limit of an infinite number of qubits. In the classical setting, there are some standard architectures which have been extensively studied (e.g. fully connected or convolutional neural networks). For such architectures, a closed formula is known for the corresponding analytic NTK. In the quantum setting, there are no standard architectures which have been proved to be more suitable for a quantum advantage yet. Therefore, as mentioned in the previous sections, we will never fix any specific architecture of the quantum circuit in our work, but we will only study the general properties holding for arbitrary overparameterized quantum neural networks. As a consequence, we will not provide any closed formula for the analytic NTK in the quantum setting. However, computing a closed formula for the analytic NTK for any finite size circuit is theoretically always possible by a direct (and inefficient) calculation, as discussed in Appendix C.

As we saw in the previous sections, $N(m)$ has to grow rapidly enough in order to ensure that a hypothesis of the form (2.77) is verified. The main limit to this request is due to barren plateaus. Since, as we will see, (2.77) is going to be generalized to

$$\lim_{m\to\infty} \frac{L^\alpha m^\beta |\mathcal{M}|^\gamma |\mathcal{N}|^\delta}{\big(N_K(m)\big)^\mu N(m)} = 0 \tag{4.28}$$

for some exponents $\alpha, \beta, \gamma, \delta$ and μ that will be specified later, we may wonder if there is an analogous problem with $N_K(m)$. The aim of next lemma is to show that – differently from $N(m)$ – $N_K(m)$ does not suffer from the problem of a potential too weak growth[12]. Furthermore, we will prove that a condition (4.28) can always be translated into the slightly stronger form (2.77).

Lemma 4.5 (Bounding the NTK normalization) *If Assumption 3.6 and Assumption 4.3 are satisfied, then*

$$\Omega(1) \leq N_K(m) < O\left(|\mathcal{N}|\right), \tag{4.29}$$

which means that there exist $c_1, c_2 > 0$ *and* $m_0 \in \mathbb{N}$ *such that, for any* $m \geq m_0$

$$c_1 \leq N_K(m) \leq c_2 |\mathcal{N}|. \tag{4.30}$$

Proof See subsection 4.7. □

Remark 4.6 Since we are interested in the case $L = O(\text{poly}(m))$, we have that

$$|\mathcal{N}| \leq |\Theta| = Lm = O(\text{poly}(m)) \quad \to \quad N_K(m) = O(\text{poly}(m)) \tag{4.31}$$

independently of the behaviour of $N(m)$. Therefore, a superpolynomial suppression of $N(m)$ in (4.28) cannot be solved by the growth of $N_K(m)$. However, due to the lower bound on $N_K(m)$, we see that the problem of barren plateaus is rooted in the normalization $N(m)$ of the function and does not involve $N_K(m)$.

Remark 4.7 As a consequence of this lemma, the estimates we will present in the following can be simplified if we set

$$\frac{1}{N_K(m)} \leq c \tag{4.32}$$

or, without loss of generality,

$$\frac{1}{N_K(m)} \leq 1 \tag{4.33}$$

by a rescaling of the normalization $N_K(m)$. This of course may weaken the upper bounds or make the hypotheses of the theorems stronger, but it allows us to present the inequalities (4.28) in the cleaner form (2.77) which does not involve the computation of the asymptotic behaviour of $N_K(m)$.

[12] Because of the form of (4.28), a too weak growth would mean $N_K(m) = o(1)$ as $m \to \infty$.

4.4 Analytic solution for the linearized model

As we will prove, the first order Taylor expansion around the values of the parameters at initialization is an approximation of the model which becomes increasingly accurate as the number of qubits is large. As in the classical case, this linear approximation allows an exact solution of the evolution equations in the case of the mean squared error. Therefore, when the number of qubits is large, the analytical solution for the linear model becomes a precious result to understand the evolution of the original model, which cannot be computed analytically. So, let us start introducing the linearized model function:

$$f^{\text{lin}}(\Theta_t^{\text{lin}}, x) = f(\Theta_0, x) + \nabla_\Theta f(\Theta_0, x)^T (\Theta_t^{\text{lin}} - \Theta_0). \tag{4.34}$$

The set of equations (4.22) for the linearized model reads

$$\begin{cases} \dot{\Theta}_t^{\text{lin}} &= -\eta \nabla_\Theta f(\Theta_0, X^T) \nabla_{f^{\text{lin}}(\Theta_t, X)} \mathcal{L}^{\text{lin}}(\Theta_t^{\text{lin}}) \\ \dfrac{d}{dt} f^{\text{lin}}(\Theta_t^{\text{lin}}, x) &= -\eta\, N_K(m)\, \hat{K}_{\Theta_0}(x, X^T) \nabla_{f^{\text{lin}}(\Theta_t^{\text{lin}}, X)} \mathcal{L}^{\text{lin}}(\Theta_t^{\text{lin}}) \end{cases}, \tag{4.35}$$

where $\mathcal{L}^{\text{lin}}$ is the loss function computed on $\mathcal{D}$ using the linearized model instead of the original model. We recall that $n = |\mathcal{D}|$. In the case of the mean squared error, the solution of (4.35) can be analytically computed. We recall that

$$\begin{aligned} \mathcal{L}(\Theta) &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \left(f(\Theta, x^{(i)}) - y^{(i)} \right)^2 \\ &= \frac{1}{2n} \| f(X) - Y \|_2^2. \end{aligned} \tag{4.36}$$

After writing the gradient of the loss function with respect to the function,

$$\nabla_{f(\Theta_t, X)} \mathcal{L}(\Theta_t) = \frac{1}{n} f(\Theta_t, X) - Y, \tag{4.37}$$

the equations of both the original and the linearized model become

$$\begin{cases} \dot{\Theta}_t &= -\dfrac{\eta}{n} \nabla_\Theta f(\Theta_t, X^T)(f(\Theta_t, X) - Y) \\ \dfrac{d}{dt} f(\Theta_t, x) &= -\eta \dfrac{N_K(m)}{n} \hat{K}_{\Theta_t}(x, X^T)\, (f(\Theta_t, X) - Y) \end{cases}, \tag{4.38}$$

$$\begin{cases} \dot{\Theta}_t^{\text{lin}} &= -\dfrac{\eta}{n} \nabla_\Theta f(\Theta_0, X^T)(f^{\text{lin}}(\Theta_t^{\text{lin}}, X) - Y) \\ \dfrac{d}{dt} f^{\text{lin}}(\Theta_t^{\text{lin}}, x) &= -\eta \dfrac{N_K(m)}{n} \hat{K}_{\Theta_0}(x, X^T) \left(f^{\text{lin}}(\Theta_t^{\text{lin}}, X) - Y \right) \end{cases}. \tag{4.39}$$

Using the notation introduced above, the equations of evolution for the model and its linearized version evalued at the dataset inputs read

$$\dot{F}(t) = -\eta \frac{N_K(m)}{n} \hat{K}_{\Theta_t} (F(t) - Y), \tag{4.40}$$

$$\dot{F}^{\text{lin}}(t) = -\eta \frac{N_K(m)}{n} \hat{K}_{\Theta_0} (F^{\text{lin}}(t) - Y), \tag{4.41}$$

where $\hat{K}_{\Theta_0} \equiv \hat{K}_{\Theta_0}(X, X^T)$ and $F^{\text{lin}}(t) = f^{\text{lin}}(\Theta_t^{\text{lin}}, X)$. In order to write the analytical solution for a new input, we assume $\hat{K}_{\Theta_0}$ to be invertible[13]. The solution of (4.41) gives the evolution of the linearized function evalued at the inputs of the dataset. By definition $F^{\text{lin}}(0) = F(0)$:

$$F^{\text{lin}}(t) = e^{-\eta \frac{N_K(m)}{n} t \hat{K}_{\Theta_0}} (F(0) - Y) + Y. \tag{4.42}$$

[13] This is ensured asymptotically by Assumption 4.3 and by the statements (4.7) and (4.24).

For a new input $x \in \mathcal{X}$

$$\begin{aligned}\frac{d}{dt} f^{\text{lin}}(\Theta_t^{\text{lin}}, x) &= -\eta \frac{N_K(m)}{n} \hat{K}_{\Theta_0}(x, X^T) \cdot (F^{\text{lin}}(t) - Y) \\ &= -\eta \frac{N_K(m)}{n} \hat{K}_{\Theta_0}(x, X^T) e^{-\eta \frac{N_K(m)}{n} \hat{K}_{\Theta_0} t} (F(0) - Y). \end{aligned} \tag{4.43}$$

Therefore

$$f^{\text{lin}}(\Theta_t^{\text{lin}}, x) = f(\Theta_0, x) - \hat{K}_{\Theta_0}(x, X^T) \hat{K}_{\Theta_0}^{-1} \left(\mathbb{1} - e^{-\eta \frac{N_K(m)}{n} \hat{K}_{\Theta_0} t} \right) (F(0) - Y). \tag{4.44}$$

Assumption 4.6 *As m increases, we fix a constant $\eta_0 > 0$ and we rescale the learning rate*

$$\eta = \frac{n}{N_K(m)} \eta_0. \tag{4.45}$$

Hence, (4.44) reads

$$f^{\text{lin}}(\Theta_t^{\text{lin}}, x) = f(\Theta_0, x) - \hat{K}_{\Theta_0}(x, X^T) \hat{K}_{\Theta_0}^{-1} \left(\mathbb{1} - e^{-\eta_0 \hat{K}_{\Theta_0} t} \right) (F(0) - Y). \tag{4.46}$$

The theorems in the following subsection show that, in the limit $m \to \infty$, $\{f^{\text{lin}}(\Theta_t^{\text{lin}}, x)\}_{x \in \mathcal{X}}$ is a Gaussian process and $f(\Theta_t, x)$ converges to $f^{\text{lin}}(\Theta_t^{\text{lin}}, x)$.

4.5 Convergence to the linearized model

The following theorem is a generalization of [38, Theorem 1] to quantum circuits which are not geometrically local and whose number of layer may depend on m.

Theorem 4.7 (NTK concentration at initialization) *Calling m the number of qubits and $L(m)$ the number of layers, let us assume that*

$$\lim_{m \to \infty} \frac{1}{N_K^2(m)} \frac{\Sigma_2 |\mathcal{M}|^2 |\mathcal{N}|^2}{N^4(m)} = 0. \tag{4.47}$$

When a parameterized quantum circuit is initialized at random, i.e., the parameters θ_i are independent random variables, the empirical NTK converges in probability to the analytic NTK as $m \to \infty$. In particular, there exists a constant c such that, for any $x, x' \in \mathcal{X}$, we have

$$\begin{aligned}\mathbb{P}\left[|\hat{K}_\Theta(x, x') - K(x, x')| \geq \epsilon \right] &\leq \exp\left[-c N_K^2(m) \frac{N^4(m)}{\Sigma_2 |\mathcal{M}|^2 |\mathcal{N}|^2} \epsilon^2 \right] \\ &\leq \exp\left[-c N_K^2(m) \frac{N^4(m)}{L m |\mathcal{M}|^4 |\mathcal{N}|^2} \epsilon^2 \right]. \end{aligned} \tag{4.48}$$

Proof See subsection 4.8. □

Remark 4.8 The hypothesis (4.47) can be simplified with the (slightly stronger) request

$$\lim_{m \to \infty} \frac{L m |\mathcal{M}|^4 |\mathcal{N}|^2}{N^4(m)} = 0, \tag{4.49}$$

which may be easier to be operationally verified, since the computation of Σ_2 and $N_K(m)$ are not required.

Denoting $\xrightarrow{p}$ the convergence in probability, and recalling that the convergence in probability implies the convergence in distribution [45], we are going to prove, informally speaking, the following implication:

$$\left.\begin{aligned} f(\Theta_0, \cdot\,) &\xrightarrow{d} f^{(\infty)}(\,\cdot\,) \\ \hat{K}_{\Theta_0}(x, x') &\xrightarrow{p} K(x, x') \to \bar{K}(x, x') \end{aligned}\right\} \quad \to \quad \begin{aligned} &\lim_{m \to \infty} f^{\text{lin}}(\Theta_t^{\text{lin}}, x) \\ &\text{can be written in terms of} \\ &f^{(\infty)}(\,\cdot\,) \text{ and } \bar{K}(\,\cdot\,, \cdot\,). \end{aligned} \tag{4.50}$$

The advantage of writing $\lim_{m\to\infty} f^{\text{lin}}(\Theta_t^{\text{lin}}, x)$ in terms of $f^{(\infty)}(\,\cdot\,)$ and $\bar{K}(\,\cdot\,,\,\cdot\,)$ is that the first object is a random variable with a known Gaussian process distribution, while the second object is a deterministic function. So, the distribution of $\lim_{m\to\infty} f^{\text{lin}}(\Theta_t, x)$ can be defined by means of the following corollaries of Lemma 4.7.

Corollary 4.8 *If (4.47) and Assumption 4.3 hold, the empirical NTK appearing in the solutions (4.46) of the equations of the evolution under gradient flow converges in probability to the limit function $\bar{K}$ of the analytical NTK. Therefore*

$$f^{\text{lin}}(\Theta_t^{\text{lin}}, x) \xrightarrow{d} f^{(\infty)}(x) - \bar{K}(x, X^T)\bar{K}^{-1}\left(\mathbb{1} - e^{-\eta_0 \bar{K} t}\right)(F^{(\infty)} - Y) \tag{4.51}$$

as $m \to \infty$. At the RHS, $f^{(\infty)}(x)$ and $F^{(\infty)} = f^{(\infty)}(X)$ are the random variables obtained in the limit $m \to \infty$ over random initialization according to Theorem 3.7.

Proof See subsection 4.9. □

Corollary 4.9 (The linearized model is a GP during all the evolution) *As $m \to \infty$, $\{f^{\text{lin}}(\Theta_t^{\text{lin}}, x)\}_{x\in\mathcal{X}}$ converges in distribution to a Gaussian process $\{f_t^{(\infty)}(x)\}_{x\in\mathcal{X}}$ with mean and covariance*

$$\mu_t(x) = \bar{K}(x, X^T)\bar{K}^{-1}\left(\mathbb{1} - e^{-\eta_0 \bar{K} t}\right) Y, \tag{4.52}$$

$$\begin{aligned}\mathcal{K}_t(x, x') = {} & \mathcal{K}_0(x, x') \\ & - \bar{K}(x, X^T)\bar{K}^{-1}\left(\mathbb{1} - e^{-\eta_0 \bar{K} t}\right)\mathcal{K}_0(X, x') \\ & - \bar{K}(x', X^T)\bar{K}^{-1}\left(\mathbb{1} - e^{-\eta_0 \bar{K} t}\right)\mathcal{K}_0(X, x) \\ & + \bar{K}(x, X^T)\bar{K}^{-1}\left(\mathbb{1} - e^{-\eta_0 \bar{K} t}\right)\mathcal{K}_0(X, X^T)\left(\mathbb{1} - e^{-\eta_0 \bar{K} t}\right)\bar{K}^{-1}K(X, x').\end{aligned} \tag{4.53}$$

Proof See subsection 4.9. □

Now we state the theorems which prove that $f(\Theta_t, x)$ and $f^{\text{lin}}(\Theta_t^{\text{lin}}, x)$ are asymptotically close as $m \to \infty$, adapting the strategy presented in [22] to the quantum setting.
First, we show that the model converges to the examples as $t \to \infty$ and that it enters the *lazy regime* as $m \to \infty$, i.e.

1. both $\sup_{t\geq 0} \|\Theta_t - \Theta_0\|_\infty$ and $\sup_{t\geq 0} \|\Theta_t^{\text{lin}} - \Theta_t\|_\infty$ converge to zero in the limit of infinitely many qubits;
2. as a consequence, the linear model becomes an increasingly good approximation of the original model during the training.

Therefore, the distribution of the trained model with many qubits will be close to the distribution of the linearized model.

Theorem 4.10 (Convergence to the examples and lazy training) *Let us assume that the hypotheses of Theorem 3.7 and of Theorem 4.7 are satisfied and that Assumption 4.3 holds. For any $\delta > 0$, there exist $\bar{m} \in \mathbb{N}$ and some constants R_0, R_1 such that, when applying gradient flow with learning rate $\eta = \frac{n}{N_K(m)}\eta_0$, for all $m \geq \bar{m}$ the following inequalities hold with probability at least $1 - \delta$ over random initialization for all $t \geq 0$:*

$$\|F(t) - Y\|_2 \leq R_0 \sqrt{n \log(2n)}\, e^{-\frac{\eta_0 \lambda_{\min}^K}{3} t}, \tag{4.54}$$

$$\|\Theta_t - \Theta_0\|_\infty \leq \frac{R_1}{\lambda_{\min}^K}\, n\sqrt{\log(2n)}\, \frac{|\mathcal{M}|}{N_K(m)N(m)}\left(1 - e^{-\frac{1}{3}\eta_0 \lambda_{\min}^K t}\right). \tag{4.55}$$

Proof See subsection 4.10. □

Remark 4.9 The (4.54) is equivalent to

$$\mathcal{L}(\Theta_t) \leq \frac{1}{2} R_0^2 \log(2n) e^{-\frac{2}{3}\eta_0 \lambda_{\min}^K t}. \tag{4.56}$$

As in the classical case, the consequence of the lazy training is that the NTK is "freezed" on its initialization value.

Corollary 4.11 (Freezing of the NTK) *Let us assume that the hypotheses of Theorem 3.7 and of Theorem 4.7 are satisfied and that Assumption 4.3 holds. For any $\delta > 0$, there exist $\bar{m} \in \mathbb{N}$ and a constant R_2 such that, when applying gradient flow with learning rate $\eta = \frac{n}{N_K(m)}\eta_0$, the following inequality holds for all $m \geq \bar{m}$ with probability at least $1 - \delta$ over random initialization:*

$$\sup_t \sup_{x,x' \in \mathcal{X}} |\hat{K}_{\Theta_0}(x, x') - \hat{K}_{\Theta_t}(x, x')| \leq \frac{R_2}{\lambda_{\min}^K} n\sqrt{\log(2n)} \frac{\Sigma_1 |\mathcal{M}|^3 |\mathcal{N}|}{N_K^2(m) N^3(m)}. \tag{4.57}$$

Proof See subsection 4.11. □

As we will prove in subsubsection 4.6.3 (see, in particular, Theorem 4.23), the lazy behaviour of the training ensures that the second order corrections becomes negligible in the limit of infinitely many qubits provided that the parameter at which the model is evaluated are the same:

$$|f(\Theta, x) - f^{\text{lin}}(\Theta, x)| \leq \frac{Lm|\mathcal{M}|^2|\mathcal{N}|}{N(m)} \|\Theta - \Theta_0\|_\infty^2. \tag{4.58}$$

This upper bound, combined with the lazy training estimate (4.55), will lay the foundations for the bound which takes into account the difference in the trajectories Θ_t and Θ_t^{lin} (see Lemma 4.13 and Theorem 4.14):

$$|f(\Theta_t, x) - f^{\text{lin}}(\Theta_t^{\text{lin}}, x)| \leq \left(\frac{C}{(\lambda_{\min}^K)^3} + C' \right) n^3 \log(2n) \frac{L^2 m^2 |\mathcal{M}|^6 |\mathcal{N}|^2}{N^5(m)} \log N(m), \tag{4.59}$$

for some C and C', with high probability.

Let us see more in detail the statements that lead to the convergence of the evolution of the original model to the evolution of the linearized model, which is a Gaussian process.

Theorem 4.12 (Uniform convergence to the linearized model) *Let us assume that the hypotheses of Theorem 3.7 and of Theorem 4.7 are satisfied and that Assumption 4.3 holds. For any $\delta > 0$, there exist a constant $R_1 > 0$ and $\bar{m} \in \mathbb{N}$ such that the following inequality holds for all $m \geq \bar{m}$ with probability at least $1 - \delta$ over random initialization:*

$$\sup_t \sup_{x \in \mathcal{X}} |f(\Theta_t, x) - f^{\text{lin}}(\Theta_t, x)| \leq \left(\frac{R_1}{\lambda_{\min}^K} \right)^2 n^2 \log(2n) \frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m) N^3(m)}. \tag{4.60}$$

Proof See subsection 4.11. □

Remark 4.10 Since the LHS of (4.60) contains a supremum over $x \in \mathcal{X}$, when

$$\lim_{m \to \infty} \frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m) N^3(m)} = 0, \tag{4.61}$$

the model converges to its linear approximation *uniformly* in x.

Now we have to take into account the fact that the equation of evolution for the parameters Θ_t and Θ_t^{lin} involve different cost functions – the original and the linearized ones – so they are in general different after $t = 0$. We therefore need a control on their discrepancy.

Lemma 4.13 (The parameters differ at the second order) *Let us assume that the hypotheses of Theorem 3.7 and of Theorem 4.7 are satisfied and that Assumption 4.3 holds. Then, for any $\delta > 0$ there exists C_1, C_2 and an integer $m_0 \in \mathbb{N}$ such that, for any $m \geq m_0$,*

$$\left\|\Theta_t - \Theta_t^{\text{lin}}\right\|_\infty \leq \left(\frac{C_1}{(\lambda_{\min}^K)^3} + C_2\right) n^3 \log(2n) \frac{Lm|\mathcal{M}|^5|\mathcal{N}|^2}{N^4(m)} \log N(m) \tag{4.62}$$

with probability at least $1 - \delta$.

Proof See subsection 4.12. □

Thanks to the previous lemma, we can now control the discrepancy between the two models taking into account the different trajectories in the parameter space.

Theorem 4.14 (The original evolution is close to the linear evolution) *Let us assume that the hypotheses of Theorem 3.7 and of Theorem 4.7 are satisfied and that Assumption 4.3 holds. Then, for any $\delta > 0$ there exists C_1, C_2 and an integer $m_0 \in \mathbb{N}$ such that, for any $m \geq m_0$*

$$|f(\Theta_t, x) - f^{\text{lin}}(\Theta_t^{\text{lin}}, x)| \leq \left(\frac{C_3}{(\lambda_{\min}^K)^3} + C_4\right) n^3 \log(2n) \frac{L^2 m^2 |\mathcal{M}|^6 |\mathcal{N}|^2}{N^5(m)} \log N(m) \tag{4.63}$$

with probability at least $1 - \delta$.

Proof See subsection 4.13. □

The following theorem is the main result of this section.

Theorem 4.15 (Quantum neural networks as Gaussian processes) *Let us assume that a circuit satisfying Assumption 4.3 and such that*

$$\lim_{m\to\infty} \frac{L^2 m^2 |\mathcal{M}|^6 |\mathcal{N}|^3}{N^5(m)} \log N(m) = 0 \tag{4.64}$$

is randomly initialized according to Assumption 3.6 and is trained using the gradient flow evolution (4.19). Then, for any $t \geq 0$, in the limit of infinitely many qubits $m \to \infty$, $\{f(\Theta_t, x)\}_{x\in\mathcal{X}}$ converges in distribution to a Gaussian process $\{f_t^{(\infty)}(x)\}_{x\in\mathcal{X}}$ with mean and covariance

$$\mu_t(x) = \bar{K}(x, X^T)\bar{K}^{-1}\left(\mathbb{1} - e^{-\eta_0 \bar{K} t}\right) Y, \tag{4.65}$$

$$\begin{aligned}
\mathcal{K}_t(x, x') = \mathcal{K}_0(x, x') \\
&- \bar{K}(x, X^T)\bar{K}^{-1}\left(\mathbb{1} - e^{-\eta_0 \bar{K} t}\right)\mathcal{K}_0(X, x') \\
&- \bar{K}(x', X^T)\bar{K}^{-1}\left(\mathbb{1} - e^{-\eta_0 \bar{K} t}\right)\mathcal{K}_0(X, x) \\
&+ \bar{K}(x, X^T)\bar{K}^{-1}\left(\mathbb{1} - e^{-\eta_0 \bar{K} t}\right)\mathcal{K}_0(X, X^T)\left(\mathbb{1} - e^{-\eta_0 \bar{K} t}\right)\bar{K}^{-1}\bar{K}(X, x').
\end{aligned} \tag{4.66}$$

Proof See subsection 4.14. □

4.6 Some useful lemmas

In this subsection we motivate and prove some lemmas which will be frequently used in many of the proof concerning the convergence of the training.

4.6.1 Loss function at initialization

The first statement is a corollary of Theorem 3.7 and consists in a bound on the mean squared loss at initialization. The role of this theorem is to fix – with high probability – the highest value of the mean squared error that the randomly initialized model could produce in the limit of infinitely many qubits. This result is not immediate, since, even though the *local* observables are bounded and, if m is fixed, also the model is bounded, the model function could in principle diverge for some inputs in the limit of

infinitely many qubits. Therefore, the loss at initialization could be arbitrarily large. Furthermore, this result lays the groundwork for a proof of an exponential convergence of the form

$$\mathcal{L}(\Theta_t) \leq Ce^{-t/\tau} \qquad \forall\, m \geq m_0 \qquad \text{for some} \qquad C, \tau, m_0 > 0. \tag{4.67}$$

Indeed, (4.67) implies an upper bound to the loss at initialization.

Corollary 4.16 *Let us assume that the hypotheses of Theorem 3.7 are satisfied. Then, for any $\delta > 0$, there are a constant $R > 0$ and integer $m_0 \in \mathbb{N}$ such that*

$$\mathbb{P}(\|F(0) - Y\|_2 \leq R) \geq 1 - \delta \qquad \forall\, m \geq m_0 \tag{4.68}$$

over random initialization, where

$$R = R_0(\delta, \max_i |y^{(i)}|, \mathcal{K})\sqrt{n \log(2n)}. \tag{4.69}$$

Remark 4.11 If $\mathcal{L}$ is the mean squared error, the claim is equivalent to

$$\forall\, m \geq m_0 \qquad \mathbb{P}\left(\mathcal{L}(\Theta_0) \leq \frac{1}{2}R_0^2\right) \geq 1 - \delta. \tag{4.70}$$

Proof We recall that $\mathcal{K}(x, x) > 0$ for any $x \in \mathcal{X}$ by Assumption 3.6. Let

$$\sigma_i^2 = \mathcal{K}(x^{(i)}, x^{(i)}), \quad \sigma = \max_{1 \leq i \leq n} \sigma_i \quad \text{and} \quad y = \max_{1 \leq i \leq n} |y^{(i)}|. \tag{4.71}$$

By Theorem 3.7, the random variables $\{f(\Theta_0, x^{(i)})\}_{i=1,\dots,n}$ converge in distribution to $\{f^{(\infty)}(x^{(i)})\}_{i=1,\dots,n}$, which are Gaussian random variables with mean zero and variance σ_i^2. Given a constant R to be fixed later, for any $i = 1, \dots, n$ we can compute

$$\begin{aligned}
&\mathbb{P}\left[|f^{(\infty)}(x^{(i)}) - y^{(i)}| \geq \frac{R}{\sqrt{n}}\right] \\
&\qquad = \mathbb{P}\left[f^{(\infty)}(x^{(i)}) \geq \frac{R}{\sqrt{n}} + y^{(i)}\right] + \mathbb{P}\left[-f^{(\infty)}(x^{(i)}) \geq \frac{R}{\sqrt{n}} - y^{(i)}\right],
\end{aligned} \tag{4.72}$$

which can be estimated using Chernoff bound [61]:

$$\begin{aligned}
&\mathbb{P}\left[|f^{(\infty)}(x^{(i)}) - y^{(i)}| \geq \frac{R}{\sqrt{n}}\right] \\
&\qquad \leq \exp\left[-\frac{1}{2\sigma_i^2}\left(\frac{R}{\sqrt{n}} + y^{(i)}\right)^2\right] + \exp\left[-\frac{1}{2\sigma_i^2}\left(\frac{R}{\sqrt{n}} - y^{(i)}\right)^2\right] \\
&\qquad \leq 2\exp\left[-\frac{1}{2\sigma_i^2}\left(\frac{R}{\sqrt{n}} - |y^{(i)}|\right)^2\right].
\end{aligned} \tag{4.73}$$

Let

$$R_0(\delta, y, \mathcal{K}) = \left(\sqrt{1 + \frac{1}{\log 2}\log\frac{2}{\delta}} + \frac{y}{\sqrt{2}\sigma\sqrt{\log 2}}\right)\sqrt{2}\sigma \tag{4.74}$$

and

$$R = R_0\sqrt{n \log(2n)}. \tag{4.75}$$

Then

$$R = \sqrt{n}\left[\sqrt{2}\sigma\sqrt{\log(2n) + \frac{\log(2n)}{\log 2}\log\frac{2}{\delta}} + y\sqrt{\frac{\log(2n)}{\log 2}}\right] \tag{4.76}$$

$$\geq \sqrt{n}\left[\sqrt{2}\sigma\sqrt{\log(2n) + \log\frac{2}{\delta}} + y\right] \tag{4.77}$$

$$\geq \sqrt{n}\left[\sqrt{2}\sigma_i\sqrt{\log\frac{4n}{\delta}} + |y^{(i)}|\right] \qquad \text{for any } i = 1, \dots, n. \tag{4.78}$$

Since

$$\frac{R}{\sqrt{n}} \geq \sqrt{2}\sigma_i\sqrt{\log\frac{4n}{\delta}} + |y^{(i)}| \geq |y^{(i)}|, \tag{4.79}$$

we can bound (4.73) as follows

$$\begin{aligned}\mathbb{P}\left[|f^{(\infty)}(x^{(i)}) - y^{(i)}| \geq \frac{R}{\sqrt{n}}\right] &\leq 2\exp\left[-\frac{1}{2\sigma_i^2}\left(\frac{R}{\sqrt{n}} - |y^{(i)}|\right)^2\right] \\ &\leq 2\exp\left[-\frac{1}{2\sigma_i^2}\left(\sqrt{2}\sigma_i\sqrt{\log\frac{4n}{\delta}}\right)^2\right] \\ &\leq 2\exp\left[-\log\frac{4n}{\delta}\right] = \frac{\delta}{2n}. \end{aligned} \tag{4.80}$$

Since the random variables $\{f(\Theta_0, x^{(i)})\}_{i=1,\dots,n}$ converge in distribution to $\{f^{(\infty)}(x^{(i)})\}_{i=1,\dots,n}$, there exists $m_0 \in \mathbb{N}$ such that

$$\mathbb{P}\left[|f(\Theta_0, x^{(i)}) - y^{(i)}| \geq \frac{R}{\sqrt{n}}\right] \leq 2\mathbb{P}\left[|f^{(\infty)}(x^{(i)}) - y^{(i)}| \geq \frac{R}{\sqrt{n}}\right] \qquad \forall\, m \geq m_0 \tag{4.81}$$

because of Lemma 3.1. Therefore,

$$\mathbb{P}\left[|f(\Theta_0, x^{(i)}) - y^{(i)}| \geq \frac{R}{\sqrt{n}}\right] \leq \frac{\delta}{n}. \tag{4.82}$$

The union bound ensures that

$$\mathbb{P}\left[|f(\Theta_0, x^{(i)}) - y^{(i)}| \leq \frac{R}{\sqrt{n}} \quad \forall i \in \{1, \dots, n\}\right] \geq 1 - \delta. \tag{4.83}$$

So, with probability at least $1 - \delta$,

$$\|F(0) - Y\|_2^2 = \sum_{i=1}^{n}\left(f(\Theta_0, x^{(i)}) - y^{(i)}\right)^2 \leq \sum_{i=1}^{n}\frac{R^2}{n} = R^2. \tag{4.84}$$

□

4.6.2 Lipschitzness of the model

The aim of this subsection is to compute the Lipschitz constant of the model, of its gradient and of the NTK. In particular, we will prove that

$$|f(\Theta, x) - f(\Theta', x)| \leq 2\frac{\Sigma_1}{N(m)}\|\Theta - \Theta'\|_\infty, \tag{4.85}$$

$$\|\nabla_\Theta f(\Theta, x) - \nabla_\Theta f(\Theta', x)\|_\infty \leq 4\,\frac{|\mathcal{M}|^2|\mathcal{N}|}{N(m)}\|\Theta - \Theta'\|_\infty, \tag{4.86}$$

$$|\hat{K}_\Theta(x, x') - \hat{K}_{\Theta_0}(x, x')| \leq 16\frac{\Sigma_1|\mathcal{M}|^2|\mathcal{N}|}{N_K(m)N^2(m)}\|\Theta - \Theta_0\|_\infty. \tag{4.87}$$

(4.85)-(4.87) imply

$$|f(\Theta, x) - f(\Theta', x)| \leq 2\frac{Lm|\mathcal{M}|}{N(m)}\|\Theta - \Theta'\|_\infty, \tag{4.88}$$

$$\|\nabla_\Theta f(\Theta, x) - \nabla_\Theta f(\Theta', x)\|_\infty \leq 4\,\frac{|\mathcal{M}|^2|\mathcal{N}|}{N(m)}\|\Theta - \Theta'\|_\infty, \tag{4.89}$$

$$|\hat{K}_\Theta(x, x') - \hat{K}_{\Theta_0}(x, x')| \leq 16\frac{Lm|\mathcal{M}|^3|\mathcal{N}|}{N^2(m)}\|\Theta - \Theta_0\|_\infty. \tag{4.90}$$

Even though it is not suitable to discuss quantum advantages and barren plateaus, the example of [38], which we already introduced in subsection 2.5, turns out to be particularly useful to describe the behavior of the lazy training in in the new light of the previous inequalities. Indeed, recalling that, in this example,

$$L \text{ is fixed} \qquad \text{and} \qquad N(m) = \sqrt{m}, \tag{4.91}$$

we have

$$|f(\Theta, x) - f(\Theta', x)| \leq c_1(L)\sqrt{m}\,\|\Theta - \Theta'\|_\infty, \tag{4.92}$$

$$\|\nabla_\Theta f(\Theta, x) - \nabla_\Theta f(\Theta', x)\|_\infty \leq \frac{c_2(L)}{\sqrt{m}} \|\Theta - \Theta'\|_\infty, \tag{4.93}$$

$$|\hat{K}_\Theta(x, x') - \hat{K}_{\Theta_0}(x, x')| \leq c_3(L) \|\Theta - \Theta_0\|_\infty. \tag{4.94}$$

The fact that the Lipschitz constant of the model diverges with m is a necessary condition for the lazy training; otherwise, any evolution yielding a small change in the parameters

$$\|\Theta_t - \Theta_0\|_\infty = o(1) \qquad \text{as} \qquad m \to \infty \tag{4.95}$$

would asymptotically be uneffective to modify the output from the initial value to the expected label according to the dataset. Also the decreasing Lipschitz constant of the gradient is meaningful: a lazy change in the parameters induces a small change in the gradient, which therefore remains closer to the one of the linearized model. Finally, from the Lipschitz constant of the kernel we understand that it is the lazy regime to ensure the freezing of the kernel to its initial value.

The starting point to prove the Lipschitzness lemmas is the following statement.

Lemma 4.17 (Uniform bounds on the derivative of the observables) *The function $\partial_{\theta_i} \partial_{\theta_j} f_k(\Theta, x)$ is continuous with respect to Θ and the following bounds hold for any $i, j \in \{1, \dots, Lm\}$:*

$$|\partial_{\theta_i} f_k(\Theta, x)| \leq 2, \tag{4.96}$$

$$|\partial_{\theta_i} \partial_{\theta_j} f_k(\Theta, x)| \leq 4, \tag{4.97}$$

$$|\partial_{\theta_i} \partial_{\theta_j} f(\Theta, x)| \leq 4 \, \frac{|\mathcal{M}_i \cap \mathcal{M}_j|}{N(m)}. \tag{4.98}$$

Proof Let $\|\cdot\|$ be the norm of the Hilbert space of the physical states, and let $\|\cdot\|_{\mathcal{L}}$ be the operator norm

$$\|O\|_{\mathcal{L}} = \sup_{\|v\| \leq 1} \|Ov\|. \tag{4.99}$$

By Assumption 2.2 $\|\mathcal{O}_k\|_{\mathcal{L}} \leq 1$ for all $k \in \{1, \dots, m\}$, so

$$|f_k(\Theta, x)| = \left| \langle 0 | U^\dagger(\Theta, x) \mathcal{O}_k U(\Theta, x) | 0 \rangle \right| \leq 1. \tag{4.100}$$

Furthermore, by Assumption 2.2, the generators of the unitaries encoding the trainable parameters have spectrum $\{-1, +1\}$. In quantum circuits like these, the derivatives of the expectation value of any observable with respect to any parameter can be written in terms of other expectation values computed for "shifted" parameters. These identities are indeed known as parameter shift rules (see (14) in [62]):

$$\partial_{\theta_i} f_k(\Theta, x) = f_k(\Theta + \Delta^{(i)}, x) - f_k(\Theta - \Delta^{(i)}, x) \qquad \Delta^{(i)}_j := \begin{cases} \frac{\pi}{4} & j = i \\ 0 & j \neq i. \end{cases} \tag{4.101}$$

Therefore, we immediately see that

$$|\partial_{\theta_i} f_k(\Theta, x)| \leq |f_k(\Theta + \Delta^{(i)}, x)| + |f_k(\Theta - \Delta^{(i)}, x)| \leq 2 \tag{4.102}$$

Similarly, using a parameter shift rule for $f_k(\Theta \pm \Delta^{(i)}, x)$ we get

$$\begin{aligned} \partial_{\theta_j} \partial_{\theta_i} f_k(\Theta, x) &= \partial_{\theta_j} f_k(\Theta + \Delta^{(i)}, x) - \partial_{\theta_j} f_k(\Theta - \Delta^{(i)}, x) \\ &= f_k(\Theta + 2\Delta^{(i)}, x) - f_k(\Theta, x) - f_k(\Theta, x) + f_k(\Theta - 2\Delta^{(i)}, x), \\ &= f_k(\Theta + 2\Delta^{(i)}, x) - 2 f_k(\Theta, x) + f_k(\Theta - 2\Delta^{(i)}, x), \end{aligned} \tag{4.103}$$

whence

$$|\partial_{\theta_j} \partial_{\theta_i} f_k(\Theta, x)| \leq |f_k(\Theta + 2\Delta^{(i)}, x)| + 2|f_k(\Theta, x)| + |f_k(\Theta - 2\Delta^{(i)}, x)| \leq 4. \tag{4.104}$$

□

We need this further result to identify the Lipschitz constants we are looking for.

Lemma 4.18 (Bounded differential on a convex implies Lipschitz) *Let $f : \mathcal{U} \to \mathbb{R}^n$ be continuously differentiable, where $\mathcal{U} \subseteq \mathbb{R}^m$ is convex. If the differential of f is bounded, i.e.*

$$K = \max_{x \in \mathcal{U}} \|df(x)\|_{\mathcal{L}} < \infty, \tag{4.105}$$

then

$$\|f(y) - f(x)\| \leq K\|y - x\| \qquad \forall\, x, y \in \mathcal{U}. \tag{4.106}$$

Proof See e.g. [63]. □

We finally need the following technical estimate.

Lemma 4.19 *The following upper bound holds:*

$$\max_{1 \leq j \leq |\Theta|} \sum_{i=1}^{|\Theta|} |\mathcal{M}_i \cap \mathcal{M}_j| \leq |\mathcal{M}|^2 |\mathcal{N}|. \tag{4.107}$$

Proof Let us define

$$S_j = \{i \in \{1, \dots, |\Theta|\} : \mathcal{M}_i \cap \mathcal{M}_j \neq \emptyset\}, \tag{4.108}$$

so that we can write

$$\sum_{i=1}^{|\Theta|} |\mathcal{M}_i \cap \mathcal{M}_j| = \sum_{i \in S_j} |\mathcal{M}_i \cap \mathcal{M}_j| \leq \sum_{i \in S_j} |\mathcal{M}_j|, \tag{4.109}$$

$$\max_{1 \leq j \leq |\Theta|} \sum_{i=1}^{|\Theta|} |\mathcal{M}_i \cap \mathcal{M}_j| \leq \left(\max_j |S_j|\right) \left(\max_j |\mathcal{M}_j|\right). \tag{4.110}$$

An estimate of the cardinality of S_j can be done by means of the following rewriting

$$\begin{aligned} S_j &= \{i \in \{1, \dots, |\Theta|\} : \mathcal{M}_i \cap \mathcal{M}_j \neq \emptyset\} && (4.111)\\ &= \{i \in \{1, \dots, |\Theta|\} : \exists k \in 1, \dots, m : i \in \mathcal{N}_k \wedge j \in \mathcal{N}_k\} \\ &= \{i \in \{1, \dots, |\Theta|\} : \exists k \in \mathcal{M}_j : i \in \mathcal{N}_k \wedge j \in \mathcal{N}_k\} \\ &= \{i \in \{1, \dots, |\Theta|\} : \exists k \in \mathcal{M}_j : i \in \mathcal{N}_k\}, && (4.112) \end{aligned}$$

$$|S_j| \leq |\mathcal{M}_j| \max_k |\mathcal{N}_k| = |\mathcal{M}_j| |\mathcal{N}|. \tag{4.113}$$

Therefore,

$$\max_{1 \leq j \leq |\Theta|} \sum_{i=1}^{|\Theta|} |\mathcal{M}_i \cap \mathcal{M}_j| \leq \left(\max_j |\mathcal{M}_j|\right)^2 |\mathcal{N}| \leq |\mathcal{M}|^2 |\mathcal{N}|. \tag{4.114}$$

□

Lemma 4.20 (Lipschitzness of the gradient) *The following inequalities hold:*

$$|\partial_{\theta_i} f(\Theta, x)| \leq 2 \frac{|\mathcal{M}_i|}{N(m)}, \tag{4.115}$$

$$\|\partial_{\theta_i} f(\Theta, X)\|_2 \leq 2\sqrt{n}\, \frac{|\mathcal{M}|}{N(m)}, \tag{4.116}$$

$$\|\nabla_\Theta f(\Theta, x)\|_1 \leq \frac{2\Sigma_1}{N(m)} \leq 2L \frac{m}{N(m)} |\mathcal{M}|, \tag{4.117}$$

$$\|\nabla_\Theta f(\Theta, x) - \nabla_\Theta f(\Theta', x)\|_\infty \leq 4\, \frac{|\mathcal{M}|^2 |\mathcal{N}|}{N(m)} \|\Theta - \Theta'\|_\infty. \tag{4.118}$$

Proof (Lemma 4.20) We will use the bounds of Lemma 4.17 and the notation introduced in the proof of Theorem 4.7:

$$\partial_{\theta_i} f(\Theta, x) = \frac{1}{N(m)} \sum_{k=1}^{m} \partial_{\theta_i} f_k(\Theta_{\mathcal{N}_k}, x) = \frac{1}{N(m)} \sum_{k \in \mathcal{M}_i} \partial_{\theta_i} f_k(\Theta, x), \tag{4.119}$$

$$|\partial_{\theta_i} f(\Theta, x)| \leq \frac{1}{N(m)} \sum_{k \in \mathcal{M}_i} |\partial_{\theta_i} f_k(\Theta, x)| \leq 2 \frac{|\mathcal{M}_i|}{N(m)}, \tag{4.120}$$

which is (4.115). Moreover,

$$\|\partial_{\theta_i} f(\Theta, X)\|_2 = \sqrt{\sum_{j=1}^{n} \left(\partial_{\theta_i} f(\Theta, x^{(j)})\right)^2} \leq 2\sqrt{n} \frac{|\mathcal{M}_i|}{N(m)} \leq 2\sqrt{n} \frac{|\mathcal{M}|}{N(m)}, \tag{4.121}$$

which proves (4.116). We can also estimate

$$\begin{aligned} \|\nabla_\Theta f(\Theta, x)\|_1 &= \sum_{i=1}^{Lm} |\partial_{\theta_i} f(\Theta, x)| \leq \frac{2}{N(m)} \sum_{i=1}^{Lm} |\mathcal{M}_i| \\ &= \frac{2\Sigma_1}{N(m)} \leq 2L \frac{m}{N(m)} |\mathcal{M}|, \end{aligned} \tag{4.122}$$

so that (4.117) is verified.
In a similar way

$$|\partial_{\theta_i} \partial_{\theta_j} f(\Theta, x)| \leq \frac{1}{N(m)} \sum_{k \in \mathcal{M}_i \cap \mathcal{M}_j} |\partial_{\theta_i} \partial_{\theta_j} f_k(\Theta, x)| \leq 4 \frac{|\mathcal{M}_i \cap \mathcal{M}_j|}{N(m)}. \tag{4.123}$$

If we fix any $x \in \mathcal{X}$ and we call

$$h_x(\Theta) = \nabla_\Theta f(\Theta, x), \tag{4.124}$$

Lemma 4.17 ensures that $h_x(\Theta)$ is a continuously differentiable function. Given any $v \in \mathbb{R}^{|\Theta|}$, the differential of $h_x(\Theta)$ (with respect to Θ) acts as

$$dh_x(\Theta) v = \sum_{i=1}^{|\Theta|} \partial_{\theta_i} \nabla_\Theta f(\Theta, x)\, v_i. \tag{4.125}$$

The operator norm of $dh_x(\Theta) : (\mathbb{R}^{|\Theta|}, \ell^\infty) \to (\mathbb{R}^{|\Theta|}, \ell^\infty)$ is given by

$$\begin{aligned} \|dh_x(\Theta)\|_{\ell^\infty \to \ell^\infty} &= \sup_{\|v\|_\infty \leq 1} \|dh(\Theta) v\|_\infty \\ &\leq \sup_{\|v\|_\infty \leq 1} \left\| \sum_{i=1}^{|\Theta|} \partial_{\theta_i} \nabla_\Theta f(\Theta, x)\, v_i \right\|_\infty \\ &= \sup_{|v_i| \leq 1} \max_{1 \leq j \leq |\Theta|} \left| \sum_{i=1}^{|\Theta|} \partial_{\theta_i} \partial_{\theta_j} f(\Theta, x)\, v_i \right| \\ &\leq \sup_{|v_i| \leq 1} \max_{1 \leq j \leq |\Theta|} \sum_{i=1}^{|\Theta|} |\partial_{\theta_i} \partial_{\theta_j} f(\Theta, x)| \\ &= \max_{1 \leq j \leq |\Theta|} \sum_{i=1}^{|\Theta|} |\partial_{\theta_i} \partial_{\theta_j} f(\Theta, x)| \\ &\leq \frac{4}{N(m)} \max_{1 \leq j \leq |\Theta|} \sum_{i=1}^{|\Theta|} |\mathcal{M}_i \cap \mathcal{M}_j|. \end{aligned} \tag{4.126}$$

Using Lemma 4.19, we have that

$$\|dh_x(\Theta)\|_{\ell^\infty \to \ell^\infty} \leq 4 \frac{|\mathcal{M}|^2 |\mathcal{N}|}{N(m)}. \tag{4.127}$$

So, the uniform bound on the convex domain $\mathscr{P}$ of Θ ensures, by Lemma 4.18, that

$$\|\nabla_\Theta f(\Theta, x) - \nabla_\Theta f(\Theta', x)\|_\infty \leq 4 \frac{|\mathcal{M}|^2 |\mathcal{N}|}{N(m)} \|\Theta - \Theta'\|_\infty \tag{4.128}$$

and this finally proves (4.118). □

Lemma 4.21 (Lipschitzness of the model) *The following inequality holds:*

$$|f(\Theta, x) - f(\Theta', x)| \leq 2\frac{\Sigma_1}{N(m)}\|\Theta - \Theta'\|_\infty. \tag{4.129}$$

Proof Using Lemma 4.18 and Lemma 4.20:

$$|f(\Theta, x) - f(\Theta', x)| \leq \max_{\Theta \in \mathscr{P}} \|df(\Theta, x)\|_{\mathcal{L}} \|\Theta - \Theta'\|_\infty, \tag{4.130}$$

where

$$\begin{aligned}\|df(\Theta, x)\|_{\mathcal{L}} &= \sup_{\|v\|_\infty \leq 1} \left| \sum_{i=1}^{Lm} \partial_{\theta_i} f(\Theta, x)\, v_i \right| \\ &\leq \|\nabla_\Theta f(\Theta, x)\|_1 \leq 2\frac{\Sigma_1}{N(m)}.\end{aligned} \tag{4.131}$$

Therefore

$$|f(\Theta, x) - f(\Theta', x)| \leq 2\frac{\Sigma_1}{N(m)}\|\Theta - \Theta'\|_\infty. \tag{4.132}$$

□

Lemma 4.22 (Lipschitzness of the NTK) *The following inequality holds:*

$$|\hat{K}_{\Theta_0}(x, x') - \hat{K}_\Theta(x, x')| \leq 16\frac{\Sigma_1 |\mathcal{M}|^2 |\mathcal{N}|}{N_K(m) N^2(m)}\|\Theta_0 - \Theta\|_\infty. \tag{4.133}$$

Proof By Lemma 4.20, we have

$$\begin{aligned}&|\hat{K}_{\Theta_0}(x, x') - \hat{K}_\Theta(x, x')| \\ &= \frac{1}{N_K(m)} |\nabla_\Theta f(\Theta_0, x) \cdot \nabla_\Theta f(\Theta_0, x') - \nabla_\Theta f(\Theta, x) \cdot \nabla_\Theta f(\Theta, x')| \\ &= \frac{1}{N_K(m)} \Big| \big(\nabla_\Theta f(\Theta_0, x) - \nabla_\Theta f(\Theta, x)\big) \cdot \nabla_\Theta f(\Theta_0, x') \\ &\qquad + \nabla_\Theta f(\Theta, x) \cdot \big(\nabla_\Theta f(\Theta_0, x') - \nabla_\Theta f(\Theta, x')\big) \Big| \\ &\leq \frac{1}{N_K(m)} \Big(\big\|\nabla_\Theta f(\Theta_0, x) - \nabla_\Theta f(\Theta, x)\big\|_\infty \big\|\nabla_\Theta f(\Theta_0, x')\big\|_1 \\ &\qquad + \big\|\nabla_\Theta f(\Theta, x)\big\|_1 \big\|\nabla_\Theta f(\Theta_0, x') - \nabla_\Theta f(\Theta, x')\big\|_\infty \Big) \\ &\leq \frac{4|\mathcal{M}|^2|\mathcal{N}|}{N_K(m) N(m)} \left(\big\|\nabla_\Theta f(\Theta_0, x')\big\|_1 + \big\|\nabla_\Theta f(\Theta, x)\big\|_1 \right) \|\Theta_0 - \Theta\|_\infty \\ &\leq 16\frac{\Sigma_1 |\mathcal{M}|^2 |\mathcal{N}|}{N_K(m) N^2(m)}\|\Theta_0 - \Theta\|_\infty.\end{aligned} \tag{4.134}$$

□

4.6.3 Original model, linearized version and lazy training

The following Theorem 4.23 claims that the difference between the original and the linear model at any input point is bounded by a diverging factor multiplied by the square of the infinity norm of the difference between the initial parameters and a new choice of the parameters. The square is crucial, since in the lazy regime $\|\Theta_t - \Theta_0\|_\infty$ is suppressed by a factor $1/N(m)$ (see Theorem 4.10), which would not be sufficient to ensure the convergence to the linearized model without such square.

Theorem 4.23 (Discrepancy between the original and the linearized model) *For any $\Theta_0 \in \mathscr{P}$ defining the linearized model, we have the following upper bound:*

$$|f(\Theta, x) - f^{\text{lin}}(\Theta, x)| \leq \frac{Lm|\mathcal{M}|^2|\mathcal{N}|}{N(m)}\|\Theta - \Theta_0\|_\infty^2. \tag{4.135}$$

In order to understand how the convergence follows from the previous inequality, we can consider again the example of [38], i.e.

$$L \text{ is fixed} \qquad \text{and} \qquad N(m) = \sqrt{m}. \tag{4.136}$$

In this case, we can bound, by Theorem 4.10,

$$\|\Theta_t - \Theta_0\|_\infty \leq \frac{c}{\sqrt{m}}, \tag{4.137}$$

so that

$$|f(\Theta_t, x) - f^{\text{lin}}(\Theta_t, x)| \leq c' \frac{m}{\sqrt{m}} \left(\frac{c}{\sqrt{m}}\right)^2 \propto \frac{1}{\sqrt{m}}. \tag{4.138}$$

Therefore, in the limit $m \to \infty$ we have uniform convergence

$$f(\Theta_t, \cdot) \to f^{\text{lin}}(\Theta_t, \cdot). \tag{4.139}$$

Let us see in the proof how a fine estimate of the second derivative using light cones ensures the result.

Proof We apply the multivariate version of Taylor's theorem with integral remainder to $f(\Theta, x)$ with respect to the parameters.

$$f(\Theta, x) = f(\Theta_0, x) + \nabla f(\Theta_0, x)^T (\Theta - \Theta_0) + \sum_{i,j=1}^{Lm} R_{ij}(\Theta, x)(\theta_i - \theta_{i,0})(\theta_j - \theta_{j,0}) \tag{4.140}$$

with

$$R_{ij}(\Theta, x) = \frac{1}{2} \int_0^1 (1-t) \partial_{\theta_i} \partial_{\theta_j} f(\Theta_0 + t(\Theta - \Theta_0), x)\, dt. \tag{4.141}$$

By Lemma 4.17,

$$|R_{ij}(\Theta, x)| \leq \frac{1}{2} \int_0^1 (1-t) \left(4 \frac{|\mathcal{M}_i \cap \mathcal{M}_j|}{N(m)}\right) dt = \frac{|\mathcal{M}_i \cap \mathcal{M}_j|}{N(m)} \tag{4.142}$$

So

$$\begin{aligned} |f(\Theta, x) - f^{\text{lin}}(\Theta, x)| &= \left| \sum_{i,j=1}^{Lm} R_{ij}(\Theta, x)(\theta_i - \theta_{i,0})(\theta_j - \theta_{j,0}) \right| \\ &\leq \left(\sum_{i,j=1}^{Lm} \frac{|\mathcal{M}_i \cap \mathcal{M}_j|}{N(m)} \right) \|\Theta - \Theta_0\|_\infty^2 \\ &\leq \frac{1}{N(m)} \left(Lm \max_{1 \leq j \leq Lm} \sum_{i=1}^{Lm} |\mathcal{M}_i \cap \mathcal{M}_j| \right) \|\Theta - \Theta_0\|_\infty^2. \end{aligned} \tag{4.143}$$

Using Lemma 4.19, we eventually estimate

$$|f(\Theta, x) - f^{\text{lin}}(\Theta, x)| \leq \frac{Lm |\mathcal{M}|^2 |\mathcal{N}|}{N(m)} \|\Theta - \Theta_0\|_\infty^2. \tag{4.144}$$

□

4.6.4 The minimum of the loss for the linear model and the convergence of the kernel

In this subsubsection we show that the linearized model, for sufficiently large m yields

$$\min_{\Theta \in \mathbb{R}^{Lm}} \mathcal{L}^{\text{lin}}(\Theta) = 0 \tag{4.145}$$

with high probability, where $\mathcal{L}^{\text{lin}}$ is the mean square error for the linearized model. If the linear model is able to perfectly fit the dataset – i.e., $\mathcal{L}^{\text{lin}}(\Theta^*) = 0$ for some $\Theta^* \in \mathbb{R}^{Lm}$ – and the training converges to the global minimum of the cost, then, in the limit of infinitely many qubits, also the original model, which is close to its linearized version in the lazy regime, will be able to perfectly fit the dataset.

Remark 4.12 In this subsubsection we will use the result of Theorem 4.7. Even though the proof of Theorem 4.7 is delayed to subsection 4.8 for a cleaner presentation, such proof will not use any result of this subsubsection.

The starting point is a convergence lemma, which will be used in different proofs.

Lemma 4.24 (Convergence of the minimum eigevalue) *Suppose that the matrix elements of a sequence of $n \times n$ symmetric matrices M_m converge in probability to the matrix elements of a positive matrix $M_\infty \succ 0$:*

$$\lim_{m\to\infty} \mathbb{P}[|(M_m)_{ij} - (M_\infty)_{ij}| \geq \epsilon] = 0 \qquad \forall\, i,j \in \{1,\dots,n\} \quad \forall\, \delta > 0. \tag{4.146}$$

Let $\lambda^M_{\min} > 0$ be the minimum eigenvalue of M_∞. Then, for any $\delta > 0$ and $\eta \in (0,1)$ there exist $m_0 \in \mathbb{N}$ such that, for all $m \geq m_0$,

$$M_n \succ \eta \lambda^M_{\min} \mathbb{1} \tag{4.147}$$

with probability at least $1-\delta$.

Proof The convergence in probability implies that

$$\begin{aligned}
\lim_{m\to\infty} \mathbb{P}[\|M_m - M_\infty\|_F \geq \epsilon] &= \lim_{m\to\infty} \mathbb{P}\left[\sum_{i,j=1}^n |(M_m)_{ij} - (M_\infty)_{ij}|^2 \geq \epsilon^2\right] \\
&\leq \lim_{m\to\infty} \mathbb{P}\left[\max_{ij} |(M_m)_{ij} - (M_\infty)_{ij}| \geq \epsilon/n\right] \\
&\leq \lim_{m\to\infty} \sum_{i,j=1}^n \mathbb{P}\left[|(M_m)_{ij} - (M_\infty)_{ij}| \geq \epsilon/n\right] = 0.
\end{aligned} \tag{4.148}$$

Therefore, for any $\delta > 0$, there exist $m_1(\epsilon) \in \mathbb{N}$ such that, for any $m \geq m_0(\epsilon)$,

$$\|M_m - M_\infty\|_F < \epsilon \tag{4.149}$$

with probability at least $1-\delta$. Let $F_m = M_\infty - M_m$. Since $\|F_m\|_F < \epsilon$, the maximum eigenvalue of $|F_m|$ is $\lambda_m < \epsilon$, so

$$F_m \preceq |F_m| \preceq \lambda_m \mathbb{1} \prec \epsilon \mathbb{1}. \tag{4.150}$$

Let $\lambda^M_{\min}$ be the minimum eigenvalue of M_∞. The previous equation implies that

$$M_\infty - M_m \prec \epsilon \mathbb{1} \quad \to \quad M_m \succ M_\infty - \epsilon \mathbb{1} \succeq (\lambda^M_{\min} - \epsilon)\mathbb{1}. \tag{4.151}$$

Since $\lambda^M_{\min}$ is positive and $\epsilon > 0$ is arbitrary, letting $\epsilon = (1-\eta)\lambda^M_{\min} > 0$ – which fixes m_0 – the claim holds. □

Then, we need a translation of Assumption 4.3 into a property of the gradient of the function at initialization which will be used to fit the dataset in the overparameterized regime.

Lemma 4.25 *If (4.47) holds, then for any $x, x' \in \mathcal{X}$, Assumption 4.3 implies the convergence in probability of $\hat{K}_{\Theta_0}(x,x')$ to $\bar{K}(x,x')$ over random initialization.*

$$\lim_{m\to\infty} \mathbb{P}\left[|\hat{K}_{\Theta_0}(x,x') - \bar{K}(x,x')| \leq \epsilon\right] = 1 \qquad \forall\, \epsilon > 0. \tag{4.152}$$

Proof Given $\epsilon > 0$, by Assumption 4.3, there exists $\bar{m} \in \mathbb{N}$ such that

$$|K(x,x') - \bar{K}(x,x')| \leq \frac{\epsilon}{2} \qquad \forall\, m \geq \bar{m} \quad \forall\, x, x' \in \mathcal{X}. \tag{4.153}$$

Therefore, $\forall\, m \geq \bar{m} \quad \forall\, x, x' \in \mathcal{X}$,

$$\begin{aligned}
|\hat{K}_{\Theta_0}(x,x') - \bar{K}(x,x')| &\leq |\hat{K}_{\Theta_0}(x,x') - K(x,x')| + |K(x,x') - \bar{K}(x,x')| \\
&\leq |\hat{K}_{\Theta_0}(x,x') - K(x,x')| + \frac{\epsilon}{2}.
\end{aligned} \tag{4.154}$$

Furthermore, by Theorem 4.7, if $m \geq \bar{m}$

$$\begin{aligned}
\mathbb{P}\left[|\hat{K}_{\Theta_0}(x,x') - \bar{K}(x,x')| \leq \epsilon\right] &\geq \mathbb{P}\left[|\hat{K}_{\Theta_0}(x,x') - K(x,x')| + \frac{\epsilon}{2} \leq \epsilon\right] \\
&= \mathbb{P}\left[|\hat{K}_{\Theta_0}(x,x') - K(x,x')| \leq \frac{\epsilon}{2}\right]
\end{aligned}$$

$$\geq 1 - \exp\left[-cN_K^2(m)\frac{N^4(m)}{|\mathcal{M}|^2|\mathcal{N}|^2\Sigma_2}\left(\frac{\epsilon}{2}\right)^2\right]. \tag{4.155}$$

Hence

$$\lim_{m\to\infty}\mathbb{P}\left[|\hat{K}_{\Theta_0}(x,x') - \bar{K}(x,x')| \leq \epsilon\right] = 1. \tag{4.156}$$

□

Lemma 4.26 *Let $n = |\mathcal{D}|$. If (4.47) and Assumption 4.3 hold, then for any $\delta > 0$, there exists $m_0 \in \mathbb{N}$ such that, for any $m \geq m_0$,*

$$\{\partial_{\theta_i} f(\Theta_0, X)\}_{1\leq i\leq Lm} \text{ is a complete set of vectors for } \mathbb{R}^n \tag{4.157}$$

with probability at least $1-\delta$ over random initialization.

Proof Since, by Lemma 4.25, as $m \to \infty$

$$\hat{K}_{\Theta_0}(X, X^T) = \frac{1}{N_K(m)}\sum_{i=1}^{Lm}\partial_{\theta_i} f(\Theta_0, X)\partial_{\theta_i} f(\Theta_0, X^T) \tag{4.158}$$

converges in probability to $\bar{K}(X, X^T)$, which is strictly positive by Assumption 4.3, there exists $m_0 \in \mathbb{N}$ such that $\hat{K}_{\Theta_0}(X, X^T)$ is strictly positive for all $m \geq m_0$ by Lemma 4.24 with high probability over random initialization. Calling

$$v_i = \frac{1}{\sqrt{N(m)}}\partial_{\theta_i} f(\Theta_0, X) \in \mathbb{R}^n, \tag{4.159}$$

we can rewrite

$$\hat{K}_{\Theta_0} \equiv \hat{K}_{\Theta_0}(X, X^T) = \sum_{i=1}^{Lm} v_i v_i^T. \tag{4.160}$$

Since

1. $\hat{K}_{\Theta_0}$ is strictly positive, it is invertible, so its range is $\mathbb{R}^n$,
2. the range of $\hat{K}_{\Theta_0}$ is $\text{Span}\{v_i\}_{1\leq i\leq Lm}$,

we conclude that

$$\text{Span}\{v_i\}_{1\leq i\leq Lm} = \mathbb{R}^n. \tag{4.161}$$

□

Now we are ready to prove the following statement.

Corollary 4.27 *For any $\delta > 0$ there exists $m_0 \in \mathbb{N}$ such that*

$$\min_{\Theta\in\mathbb{R}^{Lm}}\mathcal{L}^{\text{lin}}(\Theta) = 0 \qquad \forall\, m \geq m_0 \tag{4.162}$$

with probability at least $1-\delta$ over random initialization.

Proof Using Lemma 4.26,

$$\{\partial_{\theta_i} f(\Theta_0, X)\}_{1\leq i\leq Lm} \tag{4.163}$$

is definitely a complete set of vectors for $\mathbb{R}^n$ with high probability. So, there exists $\alpha_1, \ldots, \alpha_{Lm}$ such that $Y - f(\Theta_0, X) \in \mathbb{R}^n$ can be written as a linear combination:

$$\alpha_1\partial_{\theta_1} f(\Theta_0, X) + \cdots + \alpha_{Lm}\partial_{\theta_{Lm}} f(\Theta_0, X) = Y - f(\Theta_0, X). \tag{4.164}$$

So, choosing

$$\Theta = \Theta_0 + \begin{pmatrix}\alpha_1\\ \vdots \\ \alpha_{Lm}\end{pmatrix}, \tag{4.165}$$

we have

$$\begin{aligned}\mathcal{L}^{\text{lin}}(\Theta) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left(f^{\text{lin}}(\Theta, x^{(i)}) - y^{(i)}\right)^2\\ &= \frac{1}{2n}\|f(\Theta_0, X) + \nabla_\Theta f(\Theta_0, X)^T(\Theta - \Theta_0) - Y\|_2^2\\ &= \frac{1}{2n}\|f(\Theta_0, X) + (Y - f(\Theta_0, X)) - Y\|_2^2 = 0.\end{aligned} \tag{4.166}$$

□

4.7 Proof of Lemma 4.5

We need some preliminary statements in order to prove Lemma 4.5.

Lemma 4.28 *Assumption 4.3 implies that each diagonal term of $\bar{K} = \bar{K}(X, X^T)$ is strictly positive:*

$$\bar{K}(x,x) > 0 \qquad \forall\,(x,\,\cdot\,) \in \mathcal{D}. \tag{4.167}$$

Proof By Assumption 4.3, $\bar{K}$ is a strictly positive matrix, so its diagonal elements are strictly positive. □

Corollary 4.29 *Under Assumption 4.3, the following property of $\bar{K}$ holds:*

$$\exists\, x \in \mathcal{X} \quad \textit{such that} \quad \bar{K}(x,x) > 0 \quad \mathcal{K}(x,x) > 0. \tag{4.168}$$

Proof Fix $x \in \mathcal{X}$ such that $(x,\cdot) \in \mathcal{D}$. By Lemma 4.28, we know that $\bar{K}(x,x) > 0$. Furthermore, by Assumption 3.6, $\mathcal{K}(x,x) > 0$. □

Lemma 4.30 *Under the Assumption 4.4, the following inequalities hold for all $x \in \mathcal{X}$*

$$4\mathbb{E}\left[f^2(\Theta,x)\right] \leq \mathbb{E}\left[\|\nabla_\Theta f(\Theta,x)\|^2\right] \leq 4\,|\mathcal{N}|\,\mathbb{E}\left[f^2(\Theta,x)\right]. \tag{4.169}$$

The following statement is an improvement of [16, Lemma 1] for our case: in the proof we use stronger bounds, in order to obtain a more effective constraint to our purpose.

Proof Since $f(\Theta,x)$ is periodic with period π in each component, we can consider the Fourier series decomposition

$$f(\Theta,x) = \sum_{v\in\mathbb{Z}^{Lm}} \tilde{f}_v(x) e^{2i\Theta\cdot v} \quad \text{where} \quad \tilde{f}_v(x) = \int_0^\pi \prod_{i=1}^{Lm} \left(\frac{d\theta_i}{\pi}\right) e^{-2i\Theta\cdot v} f(\Theta,x). \tag{4.170}$$

$f(\Theta,x)$ is of the form

$$\begin{aligned} f(\Theta,x) &= \left\langle 0 \middle| U^\dagger(\Theta,x)\mathcal{M}U(\Theta,x) \middle| 0 \right\rangle \\ &= \langle 0|S_1^\dagger W_1^\dagger(\Theta) S_1 \cdots W_L^\dagger(\Theta) S_{L+1}^\dagger \mathcal{M} S_{L+1} W_L(\Theta) \cdots S_2 W_1(\Theta) S_1|0\rangle, \end{aligned} \tag{4.171}$$

where

$$W_i(\Theta) = \bigotimes_{i=1}^m e^{-i\theta_i \mathcal{G}_i}. \tag{4.172}$$

We can consider the projectors $P_i^\pm$ on the eigenspaces of $\mathcal{G}_i$:

$$\mathcal{G}_i = P_i^+ - P_i^- \quad P_i^+ + P_i^- = \mathbb{1} \tag{4.173}$$

In this way, we can rewrite

$$\begin{aligned} W_i(\theta_i) = e^{-i\theta_i\mathcal{G}_i} &= e^{-i\theta_i P_i^+} e^{i\theta_i P_i^-} \\ &= e^{-i\theta_i P_i^+}(P_i^+ + P_i^-) e^{i\theta_i P_i^-} \\ &= e^{-i\theta_i} P_i^+ + e^{i\theta_i} P_i^-. \end{aligned} \tag{4.174}$$

This means that in

$$f(\Theta,x) = \langle 0|S_1^\dagger \cdots W_i^\dagger(\theta_i) \cdots S_{L+1}^\dagger \mathcal{M} S_{L+1} \cdots W_i(\theta_i) \cdots S_1|0\rangle \tag{4.175}$$

the only dependence on θ_i will appear as

$$f(\Theta,x) = e^{-2\theta_i} f^{(i,-1)}(\Theta,x) + f^{(i,0)}(\Theta,x) + e^{+2\theta_i} f^{(i,+1)}(\Theta,x), \tag{4.176}$$

where $f^{(i,\ell)}(\Theta,x)$ does not depend on θ_i. Therefore, we know that

$$\exists i \in \{1,\dots,Lm\} : v_i \notin \{0,\pm 1\} \quad \to \quad \tilde{f}_v(x) = 0, \tag{4.177}$$

so we can restrict the sum of the Fourier expansion to

$$f(\Theta,x) = \sum_{v\in\{0,\pm1\}^{Lm}} \tilde{f}_v(x) e^{2i\Theta\cdot v}. \tag{4.178}$$

Writing $f(\Theta, x)$ as sum of local observables depending on a limited light cone, we can further restrict the set of v such that $\tilde{f}_v(x) \neq 0$. Let

$$\mathrm{supp}(v) = \{i \in \{1, \dots, Lm\} : v_i \neq 0\}. \tag{4.179}$$

Then

$$\begin{aligned}
\tilde{f}_v(x) &= \int_0^\pi \prod_{i=1}^{Lm} \left(\frac{d\theta_i}{\pi}\right) e^{-2i\Theta\cdot v} \sum_{k=1}^{m} f_k(\Theta_{\mathcal{N}_k}, x) \\
&= \sum_{k=1}^{m} \left(\int_0^\pi \prod_{i\notin\mathcal{N}_k} \left(\frac{d\theta_i}{\pi} e^{-2i\theta_i v_i}\right)\right) \left(\int_0^\pi \prod_{i\in\mathcal{N}_k} \left(\frac{d\theta_i}{\pi} e^{-2i\theta_i v_i}\right) f_k(\Theta_{\mathcal{N}_k}, x)\right) \\
&= \sum_{k=1}^{m} \chi_{\mathcal{N}_k}(\mathrm{supp}(v)) \left(\int_0^\pi \prod_{i\in\mathcal{N}_k} \left(\frac{d\theta_i}{\pi} e^{-2i\theta_i v_i}\right) f_k(\Theta_{\mathcal{N}_k}, x)\right),
\end{aligned} \tag{4.180}$$

where

$$\chi_{\mathcal{N}_k}(\mathrm{supp}(v)) = \begin{cases} 1 & \text{if } \mathrm{supp}(v) \subseteq \mathcal{N}_k \\ 0 & \text{otherwise} \end{cases}. \tag{4.181}$$

Hence,

$$\tilde{f}_v(x) \neq 0 \quad \to \quad \exists\, k \in \{1, \dots, m\} : \mathrm{supp}(v) \subseteq \mathcal{N}_k. \tag{4.182}$$

This means that

$$\tilde{f}_v(x) \neq 0 \quad \to \quad |\mathrm{supp}(v)| \leq \max_{1\leq k\leq m} |\mathcal{N}_k| = |\mathcal{N}|. \tag{4.183}$$

The function $f(\Theta, x)$ is differentiable, hence

$$\partial_{\theta_j} f(\Theta, x) = 2i \sum_{v\in\{0,\pm1\}^{Lm}} v_j \tilde{f}_v(x) e^{2i\Theta\cdot v}. \tag{4.184}$$

By Parseval's identity

$$\mathbb{E}\left[f^2(\Theta, x)\right] = \sum_{v\in\{0,\pm1\}^{Lm}} \left|\tilde{f}_v(x)\right|^2, \tag{4.185}$$

$$\mathbb{E}\left[\left(\partial_{\theta_j} f(\Theta, x)\right)^2\right] = 4 \sum_{v\in\{0,\pm1\}^{Lm}} v_j^2 \left|\tilde{f}_v(x)\right|^2. \tag{4.186}$$

(4.187)

Using the fact that in the sum of (4.186) only the terms multiplied by $v_j \in \{0, \pm1\}$ contribute to the result,

$$\mathbb{E}\left[\left(\partial_{\theta_j} f(\Theta, x)\right)^2\right] = 4 \sum_{\substack{v\in\{0,\pm1\}^{Lm} \\ v_j\neq 0}} v_j^2 \left|\tilde{f}_v(x)\right|^2 = 4 \sum_{\substack{v\in\{0,\pm1\}^{Lm} \\ v_j\neq 0}} \left|\tilde{f}_v(x)\right|^2, \tag{4.188}$$

$$\begin{aligned}
\mathbb{E}\left[\|\nabla_\Theta f(\Theta, x)\|^2\right] &= 4 \sum_{j=1}^{Lm} \sum_{\substack{v\in\{0,\pm1\}^{Lm} \\ v_j\neq 0}} \left|\tilde{f}_v(x)\right|^2 \\
&= 4 \sum_{v\in\{0,\pm1\}^{Lm}} \sum_{j:v_j\neq 0} \left|\tilde{f}_v(x)\right|^2 \\
&= 4 \sum_{v\in\{0,\pm1\}^{Lm}} \mathrm{supp}(v) \left|\tilde{f}_v(x)\right|^2 \\
&\leq 4\,|\mathcal{N}| \sum_{v\in\{0,\pm1\}^{Lm}} \left|\tilde{f}_v(x)\right|^2 = 4\,|\mathcal{N}|\,\mathbb{E}\left[f^2(\Theta, x)\right].
\end{aligned} \tag{4.189}$$

For the other inequality, we start again from (4.186) and, calling $0 = (0, \dots, 0) \in \mathbb{Z}^{Lm}$,

$$\begin{aligned}
\mathbb{E}\left[\|\nabla_\Theta f(\Theta, x)\|^2\right] &= \sum_{j=1}^{Lm} \mathbb{E}\left[\left(\partial_{\theta_j} f(\Theta, x)\right)^2\right] = 4 \sum_{v\in\mathbb{Z}^{Lm}} \left(\sum_{j=1}^{Lm} v_j^2\right) \left|\tilde{f}_v(x)\right|^2 \\
&= 4 \sum_{v\in\mathbb{Z}^{Lm}\setminus\{0\}} v^2 \left|\tilde{f}_v(x)\right|^2 \geq 4 \sum_{v\in\mathbb{Z}^{Lm}\setminus\{0\}} \left|\tilde{f}_v(x)\right|^2
\end{aligned}$$

$$= 4 \sum_{v \in \mathbb{Z}^{Lm}} \left|\tilde{f}_v(x)\right|^2 - 4\left|\tilde{f}_0(x)\right|^2. \tag{4.190}$$

Now we notice that

$$\begin{aligned}\mathbb{E}\left[f(\Theta, x)\right] &= \sum_{v \in \mathbb{Z}^{Lm}} \tilde{f}_v(x)\mathbb{E}\left[e^{2i\Theta \cdot v}\right] = \sum_{v \in \mathbb{Z}^{Lm}} \tilde{f}_v(x) \prod_{j=1}^{Lm} \mathbb{E}\left[e^{2i\theta_j v_j}\right] \\ &= \sum_{v \in \mathbb{Z}^{Lm}} \tilde{f}_v(x)\delta_{v,0} = \tilde{f}_0(x).\end{aligned} \tag{4.191}$$

By Assumption 3.6,

$$\mathbb{E}\left[f(\Theta, x)\right] = 0 \qquad \rightarrow \qquad \tilde{f}_0(x) = 0. \tag{4.192}$$

Therefore

$$\mathbb{E}\left[\|\nabla_\Theta f(\Theta, x)\|^2\right] \geq 4 \sum_{v \in \mathbb{Z}^{Lm}} \left|\tilde{f}_v(x)\right|^2 = 4\mathbb{E}\left[f^2(\Theta, x)\right]. \tag{4.193}$$

□

To prove Lemma 4.5 we are going to use Lemma 4.30 in order to bound the normalization $N_K(m)$ in terms of the normalization $N(m)$ and of the geometric quantities related to the architecture of the circuit. We will use this notation:

$$\mathcal{K}_m(x, x') = \mathbb{E}\left[f(\Theta, x)f(\Theta, x')\right]. \tag{4.194}$$

From Assumption 3.6, there exist a limit covariance function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$$\lim_{m \to \infty} \mathcal{K}_m(x, x') = \mathcal{K}(x, x'). \tag{4.195}$$

Now we are ready to prove Lemma 4.5.
As in Corollary 4.29, let $x \in \mathcal{X}$ such that both $\bar{K}(x, x)$ and $\mathcal{K}(x, x)$ are positive. We notice that

$$K(x, x) = \frac{1}{N_K(m)}\mathbb{E}\left[\|\nabla f(\Theta, x)\|^2\right] \qquad \mathcal{K}_m(x, x) = \mathbb{E}\left[f^2(\Theta, x)\right]. \tag{4.196}$$

Therefore, the results of Lemma 4.30 can be restated as

$$4\mathcal{K}_m(x, x) \leq N_K(m)K(x, x) \leq 4\left|\mathcal{N}\right|\sqrt{\mathcal{K}_m(x, x)} \tag{4.197}$$

Dividing by $K(x, x)$,

$$4\frac{\mathcal{K}_m(x, x)}{K(x, x)} \leq N_K(m) \leq 4|\mathcal{N}|\frac{\sqrt{\mathcal{K}_m(x, x)}}{K(x, x)}. \tag{4.198}$$

Since $\mathcal{K}_m \to \mathcal{K}$ and $K \to \bar{K}$ as $m \to \infty$, there exist m_0 such that, for any $m \geq m_0$,

$$\frac{1}{2}\mathcal{K}(x, x) \leq \mathcal{K}_m(x, x) \leq 2\mathcal{K}(x, x), \qquad \frac{1}{2}\bar{K}(x, x) \leq K(x, x) \leq 2\bar{K}(x, x). \tag{4.199}$$

which implies, for all $m \geq m_0$,

$$\frac{\mathcal{K}(x, x)}{\bar{K}(x, x)} \leq N_K(m) \leq \left(8\sqrt{2}\frac{\sqrt{\mathcal{K}(x, x)}}{\bar{K}(x, x)}\right)|\mathcal{N}|. \tag{4.200}$$

4.8 Proof of Theorem 4.7

We need the following lemma.

Lemma 4.31 (McDiarmid's concentration inequality [64]) *Let $X_1, \dots, X_n$ be independent random variables, each with values in $\mathbb{X}$. Let $f : \mathbb{X}^n \to \mathbb{R}$ be a mapping such that, for every $i \in \{1, \dots, n\}$ and every $(x_1, \dots, x_n), (x_1', \dots, x_n') \in \mathbb{X}^n$ that differ only in the i-th coordinate (i.e., $\forall\, j \neq i,\ x_j = x_j'$),*

$$|f(x_1, \dots, x_n) - f(x_1', \dots, x_n')| \leq c_i. \tag{4.201}$$

For any $\epsilon > 0$

$$\mathbb{P}\left(f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right). \tag{4.202}$$

Following the strategy of [38], we show that the hypotheses of McDiarmid's concentration inequality 4.31 are satistied by $\hat{K}_\Theta$.

$$\begin{aligned}
\hat{K}_\Theta(x, x') &= \frac{1}{N_K(m)} \nabla_\Theta f(\Theta, x) \cdot \nabla_\Theta f(\Theta, x') \\
&= \frac{1}{N_K(m)} \frac{1}{N^2(m)} \sum_{k,k'=1}^{m} \sum_{j=1}^{|\Theta|} \partial_{\theta_j} f_k(\Theta, x)\, \partial_{\theta_j} f_{k'}(\Theta, x') \\
&= \frac{1}{N_K(m)} \frac{1}{N^2(m)} \sum_{k,k'=1}^{m} \sum_{j \in \mathcal{N}_k \cap \mathcal{N}_{k'}} \partial_{\theta_j} f_k(\Theta, x)\, \partial_{\theta_j} f_{k'}(\Theta, x').
\end{aligned} \tag{4.203}$$

In order to clarify when a term depends or not on a particular parameter θ_i, we use the notation

$$f_k(\Theta_{\mathcal{N}_k}, x) \quad \text{instead of} \quad f_k(\Theta, x), \tag{4.204}$$

so that we emphasize that f_k depends only on θ_i with $i \in \mathcal{N}_k$.
Let $\Gamma \subseteq \{1, \dots, m\} \times \{1, \dots, m\} \times \{1, \dots, |\Theta|\}$ be the set defined as

$$\Gamma = \{(k, k', j) : j \in \mathcal{N}_k \cap \mathcal{N}_{k'}\}. \tag{4.205}$$

Let

$$T_{k,k',j}(\Theta_{\mathcal{N}_k \cup \mathcal{N}_{k'}}) = \partial_{\theta_j} f_k(\Theta_{\mathcal{N}_k}, x)\, \partial_{\theta_j} f_{k'}(\Theta_{\mathcal{N}_{k'}}, x'), \tag{4.206}$$

so that we can rewrite

$$\hat{K}_\Theta(x, x') = \frac{1}{N_K(m)} \frac{1}{N^2(m)} \sum_{(k,k',j) \in \Gamma} T_{k,k',j}(\Theta_{\mathcal{N}_k \cup \mathcal{N}_{k'}}). \tag{4.207}$$

We fix $i \in \{1, \dots, |\Theta|\}$ and we ask that $\theta_j = \theta_j'$ for all $j \neq i$. In order to compute $\hat{K}_\Theta(x, x') - \hat{K}_{\Theta'}(x, x')$, we notice that

$$i \notin \mathcal{N}_k \cup \mathcal{N}_{k'} \quad \to \quad T_{k,k',j}(\Theta_{\mathcal{N}_k \cup \mathcal{N}_{k'}}) - T_{k,k',j}(\Theta'_{\mathcal{N}_k \cup \mathcal{N}_{k'}}) = 0, \tag{4.208}$$

so we define

$$\Gamma_i = \{(k, k', j) : j \in \mathcal{N}_k \cap \mathcal{N}_{k'}, i \in \mathcal{N}_k \cup \mathcal{N}_{k'}\}. \tag{4.209}$$

Therefore, using the bound of (4.96), $|T_{k,k',j}| \leq 4$. This implies

$$\hat{K}_\Theta(x, x') - \hat{K}_{\Theta'}(x, x') = \frac{1}{N_K(m)} \frac{1}{N^2(m)} \sum_{(k,k',j) \in \Gamma_i} \left(T_{k,k',j}(\Theta) - T_{k,k',j}(\Theta')\right), \tag{4.210}$$

$$|\hat{K}_\Theta(x, x') - \hat{K}_{\Theta'}(x, x')| \leq \frac{1}{N_K(m)} \frac{1}{N^2(m)} \sum_{(k,k',j) \in \Gamma_i} \left(|T_{k,k',j}(\Theta)| + |T_{k,k',j}(\Theta')|\right)$$

$$\leq \frac{1}{N_K(m)} \frac{|\Gamma_i|}{N^2(m)} \cdot (4+4). \tag{4.211}$$

So, we need to compute $|\Gamma_i|$. A change of perspective on the constraints in the definition of Γ_i allows an easier computation:

$$\Gamma_i = \{(k, k', j) : (k \in \mathcal{M}_i \vee k' \in \mathcal{M}_i) \wedge k \in \mathcal{M}_j \wedge k' \in \mathcal{M}_j\}. \tag{4.212}$$

If we assume that the condition $k \in \mathcal{M}_i$ holds (first term in the following RHS), by symmetry we can estimate the cardinality of Γ_i:

$$\begin{aligned} \Gamma_i &= \{(k, k', j) : k \in \mathcal{M}_i \wedge k \in \mathcal{M}_j \wedge k' \in \mathcal{M}_j)\} \\ &\quad \cup \{(k, k', j) : k' \in \mathcal{M}_i \wedge k \in \mathcal{M}_j \wedge k' \in \mathcal{M}_j)\}, \end{aligned} \tag{4.213}$$

$$\begin{aligned} |\Gamma_i| &\leq 2|\{(k, k', j) : k \in \mathcal{M}_i \wedge k \in \mathcal{M}_j \wedge k' \in \mathcal{M}_j)\}| \\ &= 2|\{(k, k', j) : k \in \mathcal{M}_i \wedge j \in \mathcal{N}_k \wedge k' \in \mathcal{M}_j)\}| \\ &\leq 2|\mathcal{M}_i| \max_k |\mathcal{N}_k| \max_j |\mathcal{M}_j| \leq 2|\mathcal{M}_i||\mathcal{M}||\mathcal{N}|. \end{aligned} \tag{4.214}$$

Therefore, in McDiarmid's concentration inequality 4.31, we have

$$c_i = 16 \frac{1}{N_K(m)} \frac{|\mathcal{M}||\mathcal{N}|}{N^2(m)} |\mathcal{M}_i|. \tag{4.215}$$

Summing the squares of c_i:

$$\sum_{i=1}^{|\Theta|} c_i^2 = \left(16 \frac{|\mathcal{M}||\mathcal{N}|}{N_K(m)N^2(m)}\right)^2 \sum_{i=1}^{|\Theta|} |\mathcal{M}_i|^2 = \left(16 \frac{|\mathcal{M}||\mathcal{N}|}{N_K(m)N^2(m)}\right)^2 \Sigma_2. \tag{4.216}$$

Calling $c = 1/256$, we have

$$\begin{aligned} \mathbb{P}\left[|\hat{K}_\Theta(x, x') - K(x, x')| \geq \epsilon\right] &\leq \exp\left[-cN_K^2(m) \frac{N^4(m)}{|\mathcal{M}|^2|\mathcal{N}|^2\Sigma_2} \epsilon^2\right] \\ &\leq \exp\left[-cN_K^2(m) \frac{N^4(m)}{Lm|\mathcal{M}|^4|\mathcal{N}|^2} \epsilon^2\right]. \end{aligned} \tag{4.217}$$

4.9 Proof of Corollaries 4.8 and 4.9

We need a preliminary statement.

Theorem 4.32 (Slutsky's theorem [65]) *Let $X_1, X_2, \ldots$ be a sequence of random vectors or matrices converging in distribution to a random vector or matrix X*

$$X_k \xrightarrow{d} X \tag{4.218}$$

and let $Y_1, Y_2, \ldots$ be a sequence of random vectors or matrices converging in probability to a constant vector or matrix C

$$Y_k \xrightarrow{d} C. \tag{4.219}$$

Then

$$X_k + Y_k \xrightarrow{d} X + C, \tag{4.220}$$

$$X_k Y_k \xrightarrow{d} XC. \tag{4.221}$$

Furthermore, if C is invertible,

$$X_k / Y_k \xrightarrow{d} X/C. \tag{4.222}$$

4.9.1 Proof of Corollary 4.8

Corollary 4.8 can be proved as follows.
By Lemma 4.25,

$$\hat{K}_\Theta(x, x') \xrightarrow{p} \bar{K}(x, x') \quad \text{as} \quad m \to \infty. \tag{4.223}$$

Furthermore, by Theorem 3.7

$$f(\Theta_0, \cdot\,) \xrightarrow{d} f^{(\infty)}(\,\cdot\,) \quad \text{as} \quad m \to \infty. \tag{4.224}$$

Let $\mathcal{F} = \{x_\alpha\}_{\alpha \in A}$ be a finite family of inputs $x_\alpha \in \mathcal{X}$ containing the inputs of the dataset

$$\{x^{(i)}\}_{1 \le i \le n} \subseteq \mathcal{F} \tag{4.225}$$

Since the solution (4.46) for the output corresponding to any input $x_\beta \in \mathcal{F}$ is the following linear combination of the outputs $\{f(\Theta_0, x_\alpha)\}_{\alpha \in A}$,

$$f^{\text{lin}}(\Theta_t^{\text{lin}}, x_\beta) = f(\Theta_0, x_\beta) - \hat{K}_{\Theta_0}(x_\beta, X^T)\hat{K}_{\Theta_0}^{-1}\left(\mathbb{1} - e^{-\eta_0 \hat{K}_{\Theta_0} t}\right)(f(\Theta_0, X) - Y) \tag{4.226}$$

we can write

$$f^{\text{lin}}(\Theta_t^{\text{lin}}, x_\beta) = \sum_{\alpha \in A} M^{(t)}_{\beta\alpha}[\hat{K}_{\Theta_0}] f(\Theta_0, x_\alpha) + \left(R^{(t)}[\hat{K}_{\Theta_0}]\right)^T Y, \tag{4.227}$$

where the entries $M^{(t)}_{\beta\alpha}[\hat{K}_{\Theta_0}]$ and the components of $R^{(t)}[\hat{K}_{\Theta_0}]$ are continous functions of the elements of matrix of the empirical NTK

$$\{\hat{K}_{\Theta_0}(x_\alpha, x_{\alpha'})\}_{\alpha, \alpha' \in A}. \tag{4.228}$$

By continuity, the (finite) matrix $M^{(t)}_{\beta\alpha}[\hat{K}_{\Theta_0}]$ and the (finite) vector $R^{(t)}[\hat{K}_{\Theta_0}]$ converge in probability to $M^{(t)}_{\beta\alpha}[\bar{K}]$ and $R^{(t)}[\bar{K}]$:

$$M^{(t)}_{\beta\alpha}[\hat{K}_{\Theta_0}] \xrightarrow{p} M^{(t)}_{\beta\alpha}[\bar{K}], \qquad R^{(t)}[\hat{K}_{\Theta_0}] \xrightarrow{p} R^{(t)}[\bar{K}]. \tag{4.229}$$

By Slutsky's theorem 4.32, we conclude that

$$\{f^{\text{lin}}(\Theta_t^{\text{lin}}, x_\beta)\}_{x_\beta \in \mathcal{F}} \xrightarrow{d} \left\{\sum_{\alpha \in A} M^{(t)}_{\beta\alpha}[\bar{K}] f^{(\infty)}(x_\alpha) + \left(R^{(t)}[\bar{K}]\right)^T Y\right\}_{x_\beta \in \mathcal{F}} \quad \text{as} \quad m \to \infty, \tag{4.230}$$

i.e.

$$f^{\text{lin}}(\Theta_t^{\text{lin}}, \cdot\,)\big|_{\mathcal{F}} \xrightarrow{d} f^{(\infty)}(\,\cdot\,)\big|_{\mathcal{F}} - \bar{K}(\,\cdot\,, X^T)\big|_{\mathcal{F}}\bar{K}^{-1}\left(\mathbb{1} - e^{-\eta_0 \bar{K} t}\right)(f^{(\infty)}(X) - Y). \tag{4.231}$$

Since we assumed $\mathcal{X}$ to be finite, this is enough to prove the convergence of the distribution to the entire Gaussian process: it is sufficient to choose $\mathcal{F} = \mathcal{X}$.

Remark 4.13 In section 6 we will generalize the convergence of $\{f(\Theta_t, x)\}_{x \in \mathcal{X}}$ to a Gaussian process for the case of $\mathcal{X}$ being infinite. This will not require to prove that also $\{f^{\text{lin}}(\Theta_t, x)\}_{x \in \mathcal{X}}$ converges to a Gaussian process: we will only need the convergence of $\{f(\Theta_t, x)\}_{x \in \mathcal{F}}$ for any $\mathcal{F}$ finite set of inputs, which, as we will see, is a corollary of the convergence of the linearized model for a finite number of inputs. Therefore, the proof given above will be enough for our purposes.

Proof of Corollary 4.9

In the limit $m \to \infty$, the solution (4.51) is a linear combination of the Gaussian processes $f^{\text{lin}}(\Theta_0, x) = f(\Theta_0, x)$ and $F(0)$, so it is a Gaussian process as well, with

$$\begin{aligned}\mu_t(x) &= \mathbb{E}\left[f_t^{(\infty)}(x)\right] \\ &= \mathbb{E}\left[f^{(\infty)}(x)\right] - \bar{K}(x, X^T)\bar{K}^{-1}\left(\mathbb{1} - e^{-\eta_0 \bar{K} t}\right)\left(\mathbb{E}\left[F^{(\infty)}\right] - Y\right)\end{aligned}$$

$$= \bar{K}(x, X^T)\bar{K}^{-1}\left(\mathbb{1} - e^{-\eta_0 \bar{K} t}\right) Y, \tag{4.232}$$

$$\begin{aligned}
\mathcal{K}_t(x, x') &= \mathbb{E}\left[\left(f_t^{(\infty)}(x) - \mu_t(x)\right)\left(f_t^{(\infty)}(x') - \mu_t(x')\right)\right] \\
&= \mathbb{E}\Big[\left(f^{(\infty)}(x) - \bar{K}(x, X^T)\bar{K}^{-1}\left(\mathbb{1} - e^{-\eta_0 \bar{K} t}\right) F^{(\infty)}\right) \times \\
&\qquad \times \left(f^{(\infty)}(x') - \bar{K}(x', X^T)\bar{K}^{-1}\left(\mathbb{1} - e^{-\eta_0 \bar{K} t}\right) F^{(\infty)}\right)\Big] \\
&= \mathcal{K}_0(x, x') \\
&\quad - \bar{K}(x, X^T)\bar{K}^{-1}\left(\mathbb{1} - e^{-\eta_0 \bar{K} t}\right)\mathcal{K}_0(X, x') \\
&\quad - \bar{K}(x', X^T)\bar{K}^{-1}\left(\mathbb{1} - e^{-\eta_0 \bar{K} t}\right)\mathcal{K}_0(X, x) \\
&\quad + \bar{K}(x, X^T)\bar{K}^{-1}\left(\mathbb{1} - e^{-\eta_0 \bar{K} t}\right)\mathcal{K}_0(X, X^T)\left(\mathbb{1} - e^{-\eta_0 \bar{K} t}\right)\bar{K}^{-1} K(X, x').
\end{aligned} \tag{4.233}$$

4.10 Proof of Theorem 4.10

Because of Corollary 4.16, we can ask

$$\mathbb{P}\left(\|F(0) - Y\|_2 < R\right) \geq 1 - \frac{\delta}{2} \quad \forall\, m \geq m_0 \text{ for some } R = \sqrt{n \log(2n)} R_0. \tag{4.234}$$

In order to simplify the notation, we introduce

$$\rho(m) = \frac{6\sqrt{n} R}{N_K(m)\lambda_{\min}^K} \frac{|\mathcal{M}|}{N(m)}. \tag{4.235}$$

Let $B_r(\Theta_0) = \{\Theta : \|\Theta - \Theta_0\|_\infty < r\}$ be the ball of center Θ_0 and radius r. By Lemma 4.24, there exists $\bar{m} \in \mathbb{N}$ such that

$$\hat{K}_{\Theta_0}(X, X^T) \succ \frac{\lambda_{\min}^K}{2}\mathbb{1} \qquad \forall m \geq \bar{m} \tag{4.236}$$

with probability at least $1 - \frac{\delta}{2}$. By uniform continuity[14] of $\Theta \mapsto \hat{K}_\Theta(X, X^T)$, if $\rho(m)$ is small enough, i.e., if m is larger than $m_1 \in \mathbb{N}$[15] (we take $m_1 \geq \bar{m}$),

$$\hat{K}_\Theta(X, X^T) \succ \frac{\lambda_{\min}^K}{3}\mathbb{1} \qquad \forall\, \Theta \in B_{\rho(m)}(\Theta_0) \qquad \forall m \geq m_1. \tag{4.237}$$

Let

$$t_1 = \inf\left\{t : \|\Theta_t - \Theta_0\|_\infty \geq \rho(m)\right\}. \tag{4.238}$$

For $t \leq t_1$ we have $\hat{K}_{\Theta_t} \succ \frac{\lambda_{\min}^K}{3}\mathbb{1}$. Recalling that

$$\frac{d}{dt} f(\Theta_t, x) = -\eta_0 \hat{K}_{\Theta_t}(x, X^T) \cdot \left(F(t) - Y\right), \tag{4.239}$$

we have, with probability at least $1 - \frac{\delta}{2}$,

$$\begin{aligned}
\frac{d}{dt}\|F(t) - Y\|_2^2 &= -2\eta_0 (F(t) - Y)^T \hat{K}_{\Theta_t}(F(t) - Y) \\
&\leq -\frac{2}{3}\eta_0 \lambda_{\min}^K \|F(t) - Y\|_2^2.
\end{aligned} \tag{4.240}$$

[14] The uniform continuity is ensured by the fact that the derivatives of the entries of $\hat{K}_\Theta(X, X^T)$ with respect to the parameters are bounded and the dimension of $\hat{K}_\Theta(X, X^T)$ is fixed.

[15] By uniform continuity of $\Theta \mapsto \hat{K}_\Theta(X, X^T)$, m_1 does not depend on Θ_0.

Therefore,

$$\begin{aligned}\|F(t)-Y\|_2^2 &\le e^{-\frac{2}{3}\eta_0\lambda_{\min}^K t}\|F(0)-Y\|_2^2\\ &\le e^{-\frac{2}{3}\eta_0\lambda_{\min}^K t}R^2,\end{aligned} \tag{4.241}$$

with probability at least $1-\delta$ by the union bound with (4.234). Recalling also that

$$\dot\Theta_t = -\frac{\eta_0}{N_K(m)}\nabla_\Theta f(\Theta_t, X)\cdot(F(t)-Y) \tag{4.242}$$

and using Lemma 4.20

$$\begin{aligned}\frac{d}{dt}|\theta_i(t)-\theta_i(0)| \le \Big|\frac{d}{dt}\theta_i(t)\Big| &= \Big|\frac{\eta_0}{N_K(m)}\partial_{\theta_i} f(\Theta,X)\cdot(F(t)-Y)\Big|\\ &\le \frac{\eta_0}{N_K(m)}\|\partial_{\theta_i} f(\Theta,X)\|_2\|F(t)-Y\|_2\\ &\le \frac{\eta_0}{N_K(m)}2\sqrt{n}\,\frac{|\mathcal{M}|}{N(m)}\,Re^{-\frac{1}{3}\eta_0\lambda_{\min}^K t},\end{aligned} \tag{4.243}$$

$$\to |\theta_i(t)-\theta_i(0)| \le \frac{6\sqrt{n}R}{N_K(m)\lambda_{\min}^K}\frac{|\mathcal{M}|}{N(m)}\left(1-e^{-\frac{1}{3}\eta_0\lambda_{\min}^K t}\right) \qquad \forall\, t\le t_1, \tag{4.244}$$

$$\to \|\Theta_t-\Theta_0\|_\infty \le \rho(m)\left(1-e^{-\frac{1}{3}\eta_0\lambda_{\min}^K t}\right) \qquad \forall\, t\le t_1, \tag{4.245}$$

with probability at least $1-\delta$. If $t_1<\infty$, then

$$\|\Theta_{t_1}-\Theta_0\|_\infty \le \rho(m)\left(1-e^{-\frac{1}{3}\eta_0\lambda_{\min}^K t_1}\right) < \rho(m) \qquad \forall\, t\le t_1, \tag{4.246}$$

but this contradicts the definition (4.238) of t_1, so we must have $t_1=\infty$. This implies that (4.241) and (4.245) holds for any $t>0$ and $m\ge\bar m=\max\{m_0,m_1\}$ with probability at least $1-\delta$, so, if we set $R_1=6R_0$, we have proved (4.54) and (4.55).

4.11 Proof of Corollary 4.11 and Theorem 4.12

4.11.1 Proof of Corollary 4.11

By Lemma 4.22 and Theorem 4.10,

$$\begin{aligned}&\sup_t|\hat K_{\Theta_t}(x,x')-\hat K_{\Theta_0}(x,x')|\\ &\qquad\le 16\frac{\Sigma_1|\mathcal{M}|^2|\mathcal{N}|}{N_K(m)N^2(m)}\sup_t\|\Theta_0-\Theta\|_\infty\\ &\qquad= 96\frac{R_0}{\lambda_{\min}^K}n\sqrt{\log(2n)}\,\frac{\Sigma_1|\mathcal{M}|^3|\mathcal{N}|}{N_K^2(m)N^3(m)}\\ &\qquad= \frac{R_2}{\lambda_{\min}^K}n\sqrt{\log(2n)}\,\frac{\Sigma_1|\mathcal{M}|^3|\mathcal{N}|}{N_K^2(m)N^3(m)}, \quad \text{where} \quad R_2=96R_0,\end{aligned} \tag{4.247}$$

with probability at least $1-\delta$.

Proof of Theorem 4.12

We simply combine the discrepancy estimate of Theorem 4.23 with the lazy training result of Theorem 4.10:

$$\sup_t|f(\Theta_t,x)-f^{\rm lin}(\Theta_t,x)| \le \frac{Lm|\mathcal{M}|^2|\mathcal{N}|}{N(m)}\|\Theta_t-\Theta_0\|_\infty^2$$

$$\leq \frac{Lm|\mathcal{M}|^2|\mathcal{N}|}{N(m)} \left(\frac{R_1}{\lambda_{\min}^K} n \sqrt{\log(2n)} \frac{|\mathcal{M}|}{N_K(m) N(m)} \right)^2$$
$$\leq \left(\frac{R_1}{\lambda_{\min}^K} \right)^2 n^2 \log(2n) \frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m) N^3(m)}. \tag{4.248}$$

4.12 Proof of Lemma 4.13

In order to prove Lemma 4.13, we need the following preliminary statement.

Lemma 4.33 (Bounding the discrepancy on the examples) *For any $\delta > 0$ there exists m_0 such that, for any $m \geq m_0$*

$$||F(t) - F^{\text{lin}}(t)||_2 \leq \frac{C}{\left(\lambda_{\min}^K\right)^2} n^2 \sqrt{n} \log(2n) \frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m) N^3(m)} \left(1 - e^{-\frac{1}{3}\eta_0 \lambda_{\min}^K t}\right) \tag{4.249}$$

with probability at least $1 - \delta$.

Proof We follow the strategy of [38], using our results to improve the final bound. Let us define

$$\Delta(t) = \|F(t) - F^{\text{lin}}(t)\|_2. \tag{4.250}$$

and, recalling that

$$\frac{d}{dt} f(\Theta_t, x^{(i)}) = \hat{K}_{\Theta_t}(x^{(i)}, X^T)(f(\Theta_t, X) - Y) \quad \text{and} \quad \frac{d}{dt} f^{\text{lin}}(\Theta_t, x^{(i)}) = \hat{K}_{\Theta_0}(x^{(i)}, X^T)(f^{\text{lin}}(\Theta_t, X) - Y), \tag{4.251}$$

let us compute

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\Delta^2(t) &= \sum_{i=1}^n \frac{1}{2}\frac{d}{dt}\left(f(\Theta_t, x^{(i)}) - f^{\text{lin}}(\Theta_t^{\text{lin}}, x^{(i)})\right)^2 \\
&= \sum_{i=1}^n \left(f(\Theta_t, x^{(i)}) - f^{\text{lin}}(\Theta_t^{\text{lin}}, x^{(i)})\right)\left(\frac{d}{dt} f(\Theta_t, x^{(i)}) - \frac{d}{dt} f^{\text{lin}}(\Theta_t^{\text{lin}}, x^{(i)})\right) \\
&= -\eta_0 \sum_{i=1}^n \left(f(\Theta_t, x^{(i)}) - f^{\text{lin}}(\Theta_t^{\text{lin}}, x^{(i)})\right) \times \\
&\quad \times \left(\hat{K}_{\Theta_t}(x^{(i)}, X^T)(f(\Theta_t, X) - Y) - \hat{K}_{\Theta_0}(x^{(i)}, X^T)(f(\Theta_t^{\text{lin}}, X) - Y)\right) \\
&= -\eta_0 (F(t) - F^{\text{lin}}(t))^T \hat{K}_{\Theta_t}(F(t) - Y) + \eta_0 (F(t) - F^{\text{lin}}(t))^T \hat{K}_{\Theta_0}(F^{\text{lin}}(t) - Y) \\
&= -\eta_0 (F(t) - F^{\text{lin}}(t))^T \hat{K}_{\Theta_t}(F(t) - Y) \\
&\quad - \eta_0 (F(t) - F^{\text{lin}}(t))^T \hat{K}_{\Theta_0}(F(t) - F^{\text{lin}}(t)) \\
&\quad + \eta_0 (F(t) - F^{\text{lin}}(t))^T \hat{K}_{\Theta_0}(F(t) - Y).
\end{aligned} \tag{4.252}$$

Noticing that $\hat{K}_{\Theta_0}$ is positive semidefinite,

$$-\eta_0 (F(t) - F^{\text{lin}}(t))^T \hat{K}_{\Theta_0}(F(t) - F^{\text{lin}}(t)) \leq 0, \tag{4.253}$$

so (4.252) reads

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\Delta^2(t) &\leq -\eta_0 (F(t) - F^{\text{lin}}(t))^T \hat{K}_{\Theta_t}(F(t) - Y) + \eta_0 (F(t) - F^{\text{lin}}(t))^T \hat{K}_{\Theta_0}(F(t) - Y) \\
&= -\eta_0 (F(t) - F^{\text{lin}}(t))^T (\hat{K}_{\Theta_t} - \hat{K}_{\Theta_0})(F(t) - Y)
\end{aligned} \tag{4.254}$$

whence

$$\begin{aligned}
\left|\Delta(t)\frac{d}{dt}\Delta(t)\right| &\leq \eta_0 \|F(t) - F^{\text{lin}}(t)\|_2 \|\hat{K}_{\Theta_t} - \hat{K}_{\Theta_0}\|_{\mathcal{L}} \|F(t) - Y\|_2 \\
&= \eta_0 \Delta(t) \|\hat{K}_{\Theta_t} - \hat{K}_{\Theta_0}\|_{\mathcal{L}} \|F(t) - Y\|_2.
\end{aligned} \tag{4.255}$$

Therefore,

$$\left|\frac{d}{dt}\Delta(t)\right| \leq \eta_0 \|\hat{K}_{\Theta_t} - \hat{K}_{\Theta_0}\|_{\mathcal{L}} ||F(t) - Y||_2. \tag{4.256}$$

By (4.54)

$$\|F(t) - Y\|_2 \leq R_0 \sqrt{n \log(2n)} e^{-\frac{1}{3}\eta_0 \lambda_{\min}^K t}, \tag{4.257}$$

and by Corollary 4.11

$$\begin{aligned}\|\hat{K}_{\Theta_t} - \hat{K}_{\Theta_0}\|_{\mathcal{L}} &\leq \|\hat{K}_{\Theta_t} - \hat{K}_{\Theta_0}\|_F \\ &\leq \frac{R_2}{\lambda_{\min}^K} n^2 \sqrt{\log(2n)} \frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m) N^3(m)},\end{aligned}$$

so that

$$|\partial_t \Delta(t)| \leq \frac{C}{3} \frac{\eta_0}{\lambda_{\min}^K} n^2 \sqrt{n} \log(2n) \frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m) N^3(m)} e^{-\frac{1}{3}\eta_0 \lambda_{\min}^K t} \qquad C = 3R_0 R_2 \tag{4.258}$$

whence

$$\Delta(t) \leq \frac{C}{\left(\lambda_{\min}^K\right)^2} n^2 \sqrt{n} \log(2n) \frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m) N^3(m)} \left(1 - e^{-\frac{1}{3}\eta_0 \lambda_{\min}^K t}\right) \tag{4.259}$$

□

Now we have all the ingredients to prove Lemma 4.13. We adapt to our case the strategy of [66]. Let us compute

$$\begin{aligned}\left\|\dot{\Theta}_t - \dot{\Theta}_t^{\text{lin}}\right\|_\infty &= \frac{\eta}{n} \left\|\nabla_\Theta f(\Theta_t, X^T)(F(t) - Y) - \nabla_\Theta f(\Theta_0, X^T)(F^{\text{lin}}(t) - Y)\right\|_\infty \\ &\leq \frac{\eta}{n} \left\|\left(\nabla_\Theta f(\Theta_t, X^T) - \nabla_\Theta f(\Theta_0, X^T)\right)(F(t) - Y)\right\|_\infty \\ &\quad + \frac{\eta}{n} \left\|\nabla_\Theta f(\Theta_0, X^T)(F^{\text{lin}}(t) - F(t))\right\|_\infty \\ &\leq \frac{\eta}{n} \sup_i \|\partial_{\theta_i} f(\Theta_t, X) - \partial_{\theta_i} f(\Theta_0, X)\|_2 \|F(t) - Y\|_2 \\ &\quad + \frac{\eta}{n} \sup_i \|\partial_{\theta_i} f(\Theta_0, X)\|_2 \|F^{\text{lin}}(t) - F(t)\|_2\end{aligned} \tag{4.260}$$

Let us bound the previous expression term by term. Combining the Lipschitzness result (4.89) with the lazy training bound (4.55), and using the convergence to the examples (4.54), we control the first term:

$$\begin{aligned}&\frac{\eta}{n} \sup_i \|\partial_{\theta_i} f(\Theta_t, X) - \partial_{\theta_i} f(\Theta_0, X)\|_2 \|F(t) - Y\|_2 \\ &\qquad \leq \frac{\eta}{n} \sqrt{n}\, 4 \frac{|\mathcal{M}|^2|\mathcal{N}|}{N(m)} \|\Theta_t - \Theta_0\|_\infty \|F(t) - Y\|_2 \\ &\qquad \leq \frac{\eta}{n} \sqrt{n}\, 4 \frac{|\mathcal{M}|^2|\mathcal{N}|}{N(m)} \frac{R_1}{\lambda_{\min}^K} n \sqrt{\log(2n)} \frac{|\mathcal{M}|}{N_K(m) N(m)} \|F(t) - Y\|_2 \\ &\qquad \leq \frac{4\eta R_1}{\lambda_{\min}^K} \sqrt{n \log(2n)} \frac{|\mathcal{M}|^3|\mathcal{N}|}{N_K(m) N^2(m)} \|F(t) - Y\|_2 \\ &\qquad \leq \frac{4\eta R_1}{\lambda_{\min}^K} \sqrt{n \log(2n)} \frac{|\mathcal{M}|^3|\mathcal{N}|}{N_K(m) N^2(m)} R_0 \sqrt{n \log(2n)} e^{-\frac{1}{3}\eta_0 \lambda_{\min}^K t} \\ &\qquad = \frac{4\eta R_0 R_1}{\lambda_{\min}^K} n \log(2n) \frac{|\mathcal{M}|^3|\mathcal{N}|}{N_K(m) N^2(m)} e^{-\frac{1}{3}\eta_0 \lambda_{\min}^K t} \\ &\qquad = \frac{4\eta_0 R_0 R_1}{\lambda_{\min}^K} n^2 \log(2n) \frac{|\mathcal{M}|^3|\mathcal{N}|}{N_K^2(m) N^2(m)} e^{-\frac{1}{3}\eta_0 \lambda_{\min}^K t} =: A(t).\end{aligned} \tag{4.261}$$

Regarding the second term, we need two different estimates to be used for "small" and "large" t, as we will show soon. The first estimate is based on the Lipschitzness of the gradient (4.116) and on Lemma 4.33:

$$\begin{aligned}&\frac{\eta}{n} \sup_i \|\partial_{\theta_i} f(\Theta_0, X)\|_2 \|F^{\text{lin}}(t) - F(t)\|_2 \\ &\qquad \leq \frac{\eta}{n} 2\sqrt{n} \frac{|\mathcal{M}|}{N(m)} \|F^{\text{lin}}(t) - F(t)\|_2\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{\eta}{n} 2\sqrt{n} \frac{|\mathcal{M}|}{N(m)} \frac{C}{\left(\lambda_{\min}^K\right)^2} n^2 \sqrt{n} \log(2n) \frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m)N^3(m)} \left(1 - e^{-\frac{1}{3}\eta_0 \lambda_{\min}^K t}\right) \\
&\leq \frac{2\eta C}{\left(\lambda_{\min}^K\right)^2} n^2 \log(2n) \frac{Lm|\mathcal{M}|^5|\mathcal{N}|}{N_K^2(m)N^4(m)} \\
&= \frac{2\eta_0 C}{\left(\lambda_{\min}^K\right)^2} n^3 \log(2n) \frac{Lm|\mathcal{M}|^5|\mathcal{N}|}{N_K^3(m)N^4(m)} =: B(t).
\end{aligned}
$$

The second estimate exploits again (4.116) and the convergence to the examples for the original model (4.54); we also need to quantify the convergence to the examples for the linearized model; this immediately follows from the analytic solution (4.42) and from Theorem 4.10:

$$
\begin{aligned}
&\frac{\eta}{n} \sup_i \|\partial_{\theta_i} f(\Theta_0, X)\|_2 \|F^{\text{lin}}(t) - F(t)\|_2 \\
&\leq \frac{\eta}{n} 2\sqrt{n} \frac{|\mathcal{M}|}{N(m)} \|F^{\text{lin}}(t) - F(t)\|_2 \\
&\leq \frac{2\eta}{\sqrt{n}} \frac{|\mathcal{M}|}{N(m)} \left(\|F^{\text{lin}}(t) - Y\|_2 + \|F(t) - Y\|_2\right) \\
&\leq \frac{2\eta}{\sqrt{n}} \frac{|\mathcal{M}|}{N(m)} \left(\left\|\left(e^{-\eta_0 t \hat{K}_{\Theta_0}}(F(0) - Y) + Y\right) - Y\right\|_2 + R_0 \sqrt{n \log(2n)} e^{-\frac{1}{3}\eta_0 \lambda_{\min}^K t}\right) \\
&\leq \frac{2\eta}{\sqrt{n}} \frac{|\mathcal{M}|}{N(m)} \left(e^{-\eta_0 \lambda_{\min} t} \|F(0) - Y\|_2 + R_0 \sqrt{n \log(2n)} e^{-\frac{1}{3}\eta_0 \lambda_{\min}^K t}\right) \\
&\leq \frac{2\eta}{\sqrt{n}} \frac{|\mathcal{M}|}{N(m)} R_0 \sqrt{n \log(2n)} \left(e^{-\eta_0 \lambda_{\min} t} + e^{-\frac{1}{3}\eta_0 \lambda_{\min}^K t}\right) \\
&\leq 4R_0 \eta \sqrt{\log(2n)} \frac{|\mathcal{M}|}{N(m)} e^{-\frac{1}{3}\eta_0 \lambda_{\min}^K t} \\
&\leq 4R_0 \eta_0 \sqrt{\log(2n)} \frac{|\mathcal{M}|}{N(m)} e^{-\frac{1}{3}\eta_0 \lambda_{\min}^K t} =: C(t).
\end{aligned} \tag{4.262}
$$

Hence

$$
\left\|\dot{\Theta}_t - \dot{\Theta}_t^{\text{lin}}\right\|_\infty \leq A(t) + B(t) \quad \text{and} \quad \left\|\dot{\Theta}_t - \dot{\Theta}_t^{\text{lin}}\right\|_\infty \leq A(t) + C(t). \tag{4.263}
$$

Defining

$$
t^* = \frac{3}{\eta_0 \lambda_{\min}^K} \log N(m), \tag{4.264}
$$

we integrate

$$
\begin{aligned}
\left\|\Theta_t - \Theta_t^{\text{lin}}\right\|_\infty &\leq \int_0^\infty A(t)dt + \int_0^{t^*} B(t)dt + \int_{t^*}^\infty C(t)dt \\
&= \frac{12R_0R_1}{(\lambda_{\min}^K)^2} n^2 \log(2n) \frac{|\mathcal{M}|^3|\mathcal{N}|}{N_K^2(m)N^2(m)} \\
&\quad + \frac{2\eta_0 C}{\left(\lambda_{\min}^K\right)^2} n^3 \log(2n) \frac{Lm|\mathcal{M}|^5|\mathcal{N}|}{N_K^3(m)N^4(m)} \frac{3}{\eta_0 \lambda_{\min}^K} \log N(m) \\
&\quad + \frac{12R_0}{\lambda_{\min}^K} \sqrt{\log(2n)} \frac{|\mathcal{M}|}{N(m)} e^{-\log N(m)}
\end{aligned} \tag{4.265}
$$

Using the bound of Lemma 3.8

$$
\frac{m|\mathcal{M}||\mathcal{N}|}{N^2(m)} \geq 1, \tag{4.266}
$$

we have

$$
\left\|\Theta_t - \Theta_t^{\text{lin}}\right\|_\infty \leq \left(\frac{6C}{(\lambda_{\min}^K)^3} + \frac{12R_0R_1}{(\lambda_{\min}^K)^2} + \frac{12R_0}{\lambda_{\min}^K}\right) n^3 \log(2n) \frac{Lm|\mathcal{M}|^5|\mathcal{N}|^2}{N^4(m)} \log N(m) \tag{4.267}
$$

and with an upper bound of the form

$$ax^3 + bx^2 + cx \leq C_1 x^3 + C_2 \tag{4.268}$$

for suitable C_1 and C_2, where, in our case, $x = 1/\lambda_{\min}^K$, $a = 6C$, $b = 12R_0R_1$ and $c = 12R_0$, we conclude

$$\left\|\Theta_t - \Theta_t^{\text{lin}}\right\|_\infty \leq \left(\frac{C_1}{(\lambda_{\min}^K)^3} + C_2\right) n^3 \log(2n) \frac{Lm|\mathcal{M}|^5|\mathcal{N}|^2}{N^4(m)} \log N(m). \tag{4.269}$$

4.13 Proof of Theorem 4.14

Using the result of Theorem 4.23 and (4.117)

$$\begin{aligned}
&|f(\Theta_t, x) - f^{\text{lin}}(\Theta_t^{\text{lin}}, x)| \\
&\quad = |f(\Theta_t, x) - f(\Theta_0, x) - \nabla_\Theta f(\Theta_0, x)^T(\Theta_t^{\text{lin}} - \Theta_0)| \\
&\quad = |f(\Theta_t, x) - f(\Theta_0, x) - \nabla_\Theta f(\Theta_0, x)^T(\Theta_t - \Theta_0) - \nabla_\Theta f(\Theta_0, x)^T(\Theta_t^{\text{lin}} - \Theta_t)| \\
&\quad \leq |f(\Theta_t, x) - f^{\text{lin}}(\Theta_t, x)| + |\nabla_\Theta f(\Theta_0, x)^T(\Theta_t^{\text{lin}} - \Theta_t)| \\
&\quad \leq \frac{Lm|\mathcal{M}|^2|\mathcal{N}|}{N(m)} \|\Theta_t - \Theta_0\|_\infty^2 + \|\nabla_\Theta f(\Theta_0, x)\|_1 \|\Theta_t^{\text{lin}} - \Theta_t\|_\infty \\
&\quad \leq \frac{Lm|\mathcal{M}|^2|\mathcal{N}|}{N(m)} \|\Theta_t - \Theta_0\|_\infty^2 + 2\frac{Lm|\mathcal{M}|}{N(m)} \|\Theta_t^{\text{lin}} - \Theta_t\|_\infty.
\end{aligned} \tag{4.270}$$

Now we use Lemma 4.13 and the lazy training inequality (4.55):

$$\begin{aligned}
&|f(\Theta_t, x) - f^{\text{lin}}(\Theta_t^{\text{lin}}, x)| \\
&\quad \leq \frac{Lm|\mathcal{M}|^2|\mathcal{N}|}{N(m)} \left(\frac{R_1^2}{(\lambda_{\min}^K)^2} n^2 \log(2n) \frac{|\mathcal{M}|^2}{N_K^2(m) N^2(m)}\right) \\
&\qquad + 2\frac{Lm|\mathcal{M}|}{N(m)} \left(\frac{C_1}{(\lambda_{\min}^K)^3} + C_2\right) n^3 \log(2n) \frac{Lm|\mathcal{M}|^5|\mathcal{N}|^2}{N^4(m)} \log N(m) \\
&\quad \leq \frac{R_1^2}{(\lambda_{\min}^K)^2} n^2 \log(2n) \frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m) N^3(m)} \\
&\qquad + 2\left(\frac{C_1}{(\lambda_{\min}^K)^3} + C_2\right) n^3 \log(2n) \frac{L^2m^2|\mathcal{M}|^6|\mathcal{N}|^2}{N^5(m)} \log N(m).
\end{aligned} \tag{4.271}$$

Therefore, is easy to show that

$$|f(\Theta_t, x) - f^{\text{lin}}(\Theta_t^{\text{lin}}, x)| \leq \left(\frac{C_3}{(\lambda_{\min}^K)^3} + C_4\right) n^3 \log(2n) \frac{L^2m^2|\mathcal{M}|^6|\mathcal{N}|^2}{N^5(m)} \log N(m). \tag{4.272}$$

4.14 Proof of Theorem 4.15

Lemma 4.34 *The hypothesis*

$$\lim_{m\to\infty} \frac{L^2m^2|\mathcal{M}|^6|\mathcal{N}|^3}{N^5(m)} \log N(m) = 0, \tag{4.273}$$

of Theorem 4.15 ensures that the hypotheses of Theorem 3.7 and of Theorem 4.7 are satisfied, i.e.

$$\lim_{m\to\infty} \frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3(m)} = 0 \qquad \textit{and} \qquad \lim_{m\to\infty} \frac{1}{N_K^2(m)} \frac{\Sigma_2|\mathcal{M}|^2|\mathcal{N}|^2}{N^4(m)} = 0. \tag{4.274}$$

Furthermore,

$$\lim_{m\to\infty} \frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m)N^3(m)} = 0, \tag{4.275}$$

i.e., the bound of Theorem 4.23 is nontrivial in the limit $m \to \infty$.

Proof We will use that $|\mathcal{M}|, |\mathcal{N}|$, and L are all bounded from below by 1, and $N_K(m)$ is asymptotically bounded by (4.30) of Lemma 4.5, and we will use the inequality of Lemma 3.8 in the form

$$1 \leq c\frac{m|\mathcal{M}||\mathcal{N}|}{N^2(m)} \tag{4.276}$$

so that

$$0 \leq \lim_{m\to\infty} \frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3(m)} \leq c \lim_{m\to\infty} \frac{m^2|\mathcal{M}|^3|\mathcal{N}|^3}{N^5(m)} \leq c \lim_{m\to\infty} \frac{L^2m^2|\mathcal{M}|^6|\mathcal{N}|^3}{N^5(m)} \log N(m) = 0, \tag{4.277}$$

$$0 \leq \lim_{m\to\infty} \frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m)N^3(m)} \leq c \lim_{m\to\infty} \frac{Lm^2|\mathcal{M}|^5|\mathcal{N}|^2}{N_K^2(m)N^5(m)} \leq c \lim_{m\to\infty} \frac{L^2m^2|\mathcal{M}|^6|\mathcal{N}|^3}{N^5(m)} \log N(m) = 0. \tag{4.278}$$

Now, noticing that (4.273) implies that $\lim_{m\to\infty} N(m) = \infty$, we conclude that

$$\begin{aligned} 0 \leq \lim_{m\to\infty} \frac{1}{N_K^2(m)} \frac{\Sigma_2|\mathcal{M}|^2|\mathcal{N}|^2}{N^4(m)} &\leq c \lim_{m\to\infty} \frac{1}{N_K^2(m)} \frac{\Sigma_2 m|\mathcal{M}|^3|\mathcal{N}|^3}{N^6(m)} \\ &\leq c \lim_{m\to\infty} \frac{1}{N(m)} \frac{L^2m^2|\mathcal{M}|^6|\mathcal{N}|^3}{N^5(m)} = 0. \end{aligned} \tag{4.279}$$

□

Now we are ready to prove Theorem 4.15.
Let $\bar{X} \in \mathbb{R}^N$ be any (finite dimensional) vector on N distinct inputs $\{\bar{x}_1, \ldots, \bar{x}_n\} \in \mathcal{X}$ and let $\Delta_t(\bar{X})$ be the random vector

$$\Delta_t(\bar{X}) = f(\Theta_t, \bar{X}) - f^{\text{lin}}(\Theta_t, \bar{X}). \tag{4.280}$$

We show that $\Delta_t(\bar{X}) \xrightarrow{p} 0$ as $m \to \infty$. Let $\epsilon, \delta > 0$. By Theorem 4.14, there exists m_0 such that, for any $m \geq m_0$,

$$\begin{aligned} \mathbb{P}\Bigg(&\sup_{t'} \sup_{x\in\mathcal{X}} |f(\Theta_{t'}, x) - f^{\text{lin}}(\Theta_{t'}^{\text{lin}}, x)| \\ &\leq \left(\frac{C_3}{(\lambda_{\min}^K)^3} + C_4\right) n^3 \log(2n) \frac{L^2m^2|\mathcal{M}|^6|\mathcal{N}|^2}{N^5(m)} \log N(m)\Bigg) \geq 1 - \delta. \end{aligned} \tag{4.281}$$

whence

$$\begin{aligned} \mathbb{P}\Bigg(&\|f(\Theta_t, \bar{X}) - f^{\text{lin}}(\Theta_t^{\text{lin}}, \bar{X})\|_2 \\ &\leq \sqrt{N}\left(\frac{C_3}{(\lambda_{\min}^K)^3} + C_4\right) n^3 \log(2n) \frac{L^2m^2|\mathcal{M}|^6|\mathcal{N}|^2}{N^5(m)} \log N(m)\Bigg) \geq 1 - \delta. \end{aligned} \tag{4.282}$$

There is also $m_1 \in \mathbb{N}$ such that

$$\sqrt{N}\left(\frac{C_3}{(\lambda_{\min}^K)^3} + C_4\right) n^3 \log(2n) \frac{L^2m^2|\mathcal{M}|^6|\mathcal{N}|^2}{N^5(m)} \log N(m) < \epsilon \qquad \forall\, m \geq m_1. \tag{4.283}$$

So, for any $m \geq \max\{m_0, m_1\}$,

$$\mathbb{P}\left(\|\Delta_t(\bar{X})\|_2 < \epsilon\right) \geq 1 - \delta, \tag{4.284}$$

so

$$\Delta_t(\bar{X}) \xrightarrow{p} 0 \quad \text{as} \quad m \to \infty. \tag{4.285}$$

By Corollary 4.9, $\{f^{\text{lin}}(\Theta_t^{\text{lin}}, x)\}_{x\in\{\bar{x}_1,\ldots,\bar{x}_n\}}$ converges in distribution to a Gaussian process with mean μ_t and covariance $\mathcal{K}_t$ defined in the statement of the corollary:

$$\{f^{\text{lin}}(\Theta_t, x)\}_{x\in\mathcal{X}} \xrightarrow{d} \{f_t^{(\infty)}(x)\}_{x\in\mathcal{X}}. \tag{4.286}$$

Now, applying Slutsky's theorem 4.32 we have

$$f(\Theta_t, \bar{X}) = f^{\text{lin}}(\Theta_t, \bar{X}) + \Delta_t(\bar{X}) \xrightarrow{d} f_t^{(\infty)}(x) \quad \text{as} \quad m \to \infty. \tag{4.287}$$

Since $\mathcal{X}$ is finite, we can choose $\bar{X} = \mathrm{vec}(\mathcal{X})$ so that we have the convergence of the full multivariate Gaussian distribution, which is the Gaussian process itself.

Remark 4.14 This argument is not valid in the case of $\mathcal{X}$ infinite because we cannot consider $\bar{X} = \mathrm{vec}(\mathcal{X})$. See section 6 for the proof in the most general case.

5 Noisy training of quantum neural networks

5.1 Gradient descent with a finite number of measurements

The model function $f(\Theta, x)$ is defined in terms of expectation values of quantum observables, whose result is not deterministic. Therefore, the algorithm has to be iterated a number of times which increases with the precision required.
Also the gradient of $f(\Theta, x)$, which is involved in the training of the neural network, has to be estimated through repeated measurements. This can be done using the parameter-shift rules. The limited precision of the model function and of its gradients has to be taken into account when we look for the convergence of the training. As depicted in Figure 20, if the variance of the estimator of the gradient of $\mathcal{L}$ is not sufficiently small, the training could converge to a neighborhood of the minimum. For instance, if the variance of the estimator is fixed, when the parameters are sufficiently close to a minimum the exact gradient becomes of the same order of the statistical noise and the evolution is subject to relevant deviations. If, at a certain point, the noise becomes much larger than the exact gradient, the parameters evolve in a random direction, possibly away from the minimum. The size of the neighborhood depends on the variance of the estimator and on the shape of the loss near the minimum.
The procedure of estimating the model function, its gradient and of computing the evolution of the parameters is done in discrete time steps. This is why we need to move from the gradient flow to the gradient descent.

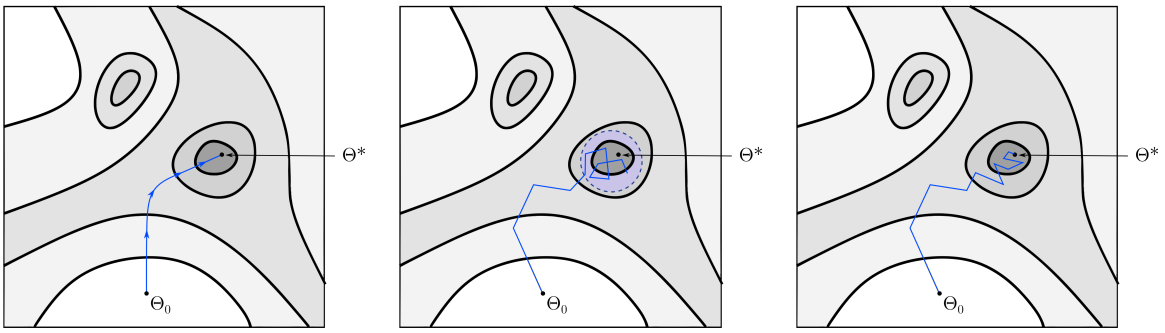


Fig. 20 Three possible scenarios: on the left, the gradient flow; in the middle, the stochastic gradient descent which converges to a neighborhood of the global minimum due to the high variance of the estimator of the gradient; on the right, the stochastic gradient descent with sufficiently small variance.

The aim of this section is to prove that for a sufficiently high number of measurement – which scales *polynomially* with the number of qubits – the training converges exponentially fast to the global minimum of the loss function, as we saw in the simpler case discussed in section 4. The strategy is based on two steps:

1. prove that the deterministic gradient descent converges to the global minimum;
2. show that, for a sufficiently small variance of the estimator of the gradient of the loss, the same convergence can be proved.

Since also in this case we retrieve the lazy training regime in the limit of infinitely many qubits, the trained model function is still a Gaussian process after random initialization and stochastic training.

The outline of this section is the following.

1. In subsection 5.2, we introduce the deterministic gradient descent and we prove that it makes the empirical loss converge to zero (Theorem 5.1).

2. In subsection 5.3, we discuss the stochastic gradient descent and we prove that the convergence of the empirical loss to zero, the lazy training and the convergence to the linearized model are still valid (Theorem 5.4). As a consequence, the model is a Gaussian process in the limit of infinitely many qubits also in the case with noise (Theorem 5.9). These two theorems are the most important results of the section and of the article.
3. In subsection 5.4, subsection 5.5 and subsection 5.8 we prove the main theorems of the previous subsections.

5.2 Deterministic gradient descent

The gradient descent equation is

$$\Theta_{t+1} - \Theta_t = -\eta \nabla_\Theta \mathcal{L}(\Theta_t). \tag{5.1}$$

As in the case of gradient flow, we consider the mean squared error function:

$$\begin{aligned}\mathcal{L}(\Theta_t) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left(f(\Theta_t, x^{(i)}) - y^{(i)}\right)^2 \\ &= \frac{1}{2n}\|F(t) - Y\|_2^2.\end{aligned} \tag{5.2}$$

This means that the equation (5.1) can be rewritten as

$$\Theta_{t+1} - \Theta_t = -\frac{\eta}{n}\nabla_\Theta F(t) \cdot (F(t) - Y). \tag{5.3}$$

We will see, in the proof of Theorem 5.9, that the solution of (5.3) for the linearized model gives, for $t \to \infty$, the same result obtained in the continuous time setting:

$$\lim_{t\to\infty} f^{\text{lin}}(\Theta_t, x) = f(\Theta_0, x) - \hat{K}_{\Theta_0}(x, X^T)\hat{K}_{\Theta_0}^{-1}(f(\Theta_0, X) - Y), \tag{5.4}$$

provided that

$$\eta_0 < \frac{2}{\lambda_{\min}^K + \lambda_{\max}^K} \tag{5.5}$$

and m is large enough (see subsection 5.8). The equation (5.4) shows that linearized model trained for infinite time perfectly reproduces the training labels.

Theorem 5.1 (Deterministic gradient descent) *Let us assume that the hypotheses of Theorem 3.7 and of Theorem 4.7 are satisfied and that Assumption 4.3 holds. Let us furthermore assume that*

$$\lim_{m\to\infty} \frac{\Sigma_1 |\mathcal{M}|^3 |\mathcal{N}|}{N_K^2(m) N^3(m)} = 0. \tag{5.6}$$

For any $\delta > 0$, there exist $\bar{m} \in \mathbb{N}$ and some constants R_0, R_1, R_2 such that, when applying gradient descent with learning rate $\eta = \frac{n}{N_K(m)}\eta_0$, where

$$\eta_0 < \frac{2}{\lambda_{\min}^K + \lambda_{\max}^K}, \tag{5.7}$$

the following inequalities hold for all $m \geq \bar{m}$ with probability at least $1 - \delta$ over random initialization:

$$\|F(t) - Y\|_2 \leq \sqrt{n\log(2n)} R_0 \left(1 - \frac{\eta_0 \lambda_{\min}^K}{3}\right)^t, \tag{5.8}$$

$$\|\Theta_t - \Theta_0\|_\infty \leq \frac{6R_0}{\lambda_{\min}^K} n\sqrt{\log(2n)} \frac{|\mathcal{M}|}{N_K(m) N(m)}. \tag{5.9}$$

Proof See subsection 5.4. □

Remark 5.1 The (5.8) is equivalent to

$$\mathcal{L}(\Theta_t) \leq \frac{R_0^2}{2} \log(2n) \left(1 - \frac{\eta_0 \lambda_{\min}^K}{3}\right)^{2t}. \tag{5.10}$$

Analogously to the continous-time case we have uniform convergence to the linearized model.

Theorem 5.2 *Let us assume that the hypotheses of Theorem 3.7 and of Theorem 5.1 are satisfied and that Assumption 4.3 holds. For any $\delta > 0$, there exist a constant $R_1 > 0$ and $\bar{m} \in \mathbb{N}$ such that the following inequality holds for all $m \geq \bar{m}$ with probability at least $1 - \delta$ over random initialization:*

$$\sup_t \sup_{x \in \mathcal{X}} |f(\Theta_t, x) - f^{\text{lin}}(\Theta_t, x)| \leq \left(\frac{6R_2}{\lambda_{\min}^K}\right)^2 n^2 \log(2n) \frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m) N^3(m)}. \tag{5.11}$$

Proof Combining Theorem 4.23 with the inequality (5.9), the bound (5.11) follows immediately. □

Also the proof of the convergence to a Gaussian process is similar to the continuous time setting, with some differences dues to the discrete nature of time in the evolution of the model. All the details of the proof of the convergence will be given for the noisy case in the following subsection and can be immediately adapted to the case discussed in this subsection by removing the additional terms due to the statistical noise.

5.3 Stochastic gradient descent

Assumption 5.3 (Unbiased gradient descent) *We consider the gradient-descent evolution*

$$\Theta_{t+1} - \Theta_t = -\eta g^{(t)}(\Theta_t), \tag{5.12}$$

where for any training time $t \in \mathbb{N}$, $g^{(t)}(\Theta_t)$ is an unbiased estimator of the gradient of the loss function in the stochastic gradient descent equation, i.e., an estimator whose average conditioned on the past history coincides with the exact value of the gradient. More formally, each $g^{(t)}(\Theta_t)$ is a random vector such that, for any $k < t$, any $\xi_k \in \mathbb{R}$ and any $\Theta_k \in \mathbb{R}^{|\Theta|}$ we have

$$\mathbb{E}\left[g_i^{(t)}(\Theta_t) \;\middle|\; g^{(k)}(\Theta_k) = \xi_k \quad \forall\, k < t\right] = \partial_{\theta_i} \mathcal{L}(\Theta) \qquad \forall\, \Theta \in \mathscr{P}. \tag{5.13}$$

Remark 5.2 The meaning of (5.13) is the following: once we are at time t, we perform the measurements of the output of circuit and we build an the estimator $g^{(t)}(\Theta_t)$ such that its expectation value is $\nabla_\Theta \mathcal{L}(\Theta)$ whatever the output of the measurements employed to determine $\{g^{(0)}, g^{(1)}, \ldots, g^{(t-1)}\}$ was. This is what we physically ask.

The aim of the following theorem is to show that, with a bound on the variance which is polynomial in the number of qubits, an unbiased estimator of the gradient of the MSE loss is suitable to achieve an exponential convergence to the global minimum with a lazy evolution of the parameters.
The idea is to compare the stochastic evolution with the exact one. In Figure 21, a tentative strategy is depicted on the left. Given an initial vector of parameters Θ_0, the stochastic evolution of the parameter is represented by the violet curve (the parameter vectors are called Θ_t), while the deterministic evolution with the exact gradient is the blue curve (the parameter vectors are called $\tilde{\Theta}_t$), which is known to converge to the global minimum by Theorem 5.1. We could try to bound the distance between the two curves using some arguments based on the triangle inequality. However, some problems may arise.

1. At each time step $t > 0$, the point Θ_t at which the estimator of the gradient is computed is different from the point $\tilde{\Theta}_t$ at which the exact gradient is computed; this means that, even if the exact evolution is close to a minimum ($\nabla_\Theta \mathcal{L}(\tilde{\Theta}_t) \leq \epsilon$), the estimated distance $\|\Theta_t - \tilde{\Theta}_t\|_\infty$ would affect the upper bound on $\nabla_\Theta \mathcal{L}(\Theta_t)$ with a bias due to a rough triangle inequality argument, i.e.

$$\nabla_\Theta \mathcal{L}(\Theta_t) \leq \nabla_\Theta \mathcal{L}(\tilde{\Theta}_t) + g(\|\Theta_t - \tilde{\Theta}_t\|_\infty). \tag{5.14}$$

   It is easy to realize that g is non decreasing. The hope is to show that somehow $\|\Theta_t - \tilde{\Theta}_t\|_\infty$ becomes small with t.

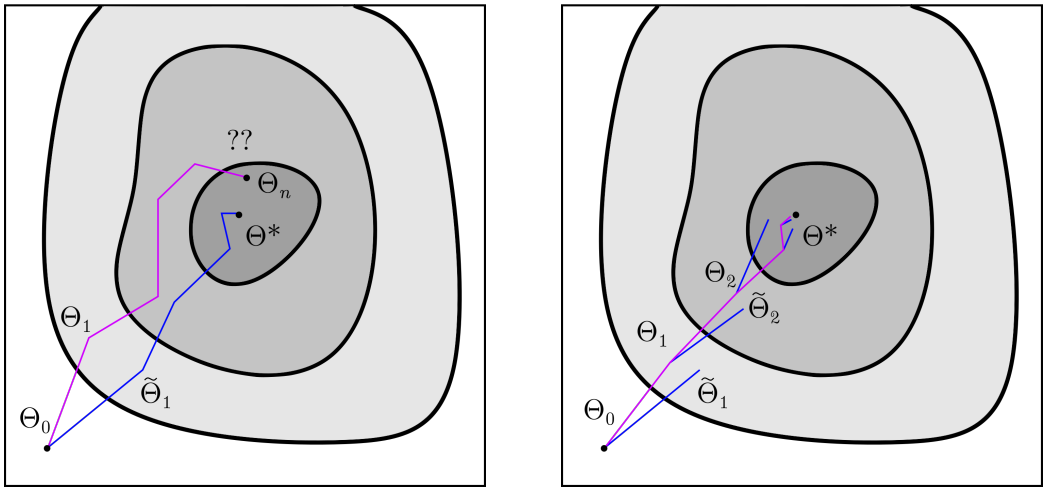


Fig. 21 The global and the local strategies to approach the stochastic gradient descent convergence. With a global strategy we lose track of the path in the parameter space.

2. However, since (5.14) asymptotically gives few information on the convergence of Θ_t to Θ^* due to the bias g, we can only worsen the estimate:

$$\|\Theta_{t+1} - \tilde{\Theta}_{t+1}\|_\infty \leq \|\Theta_{t+1} - \Theta_t\|_\infty + \|\Theta_t - \tilde{\Theta}_t\|_\infty + \|\tilde{\Theta}_{t+1} - \tilde{\Theta}_t\|_\infty. \tag{5.15}$$

   So, the bias $g(\|\Theta_t - \tilde{\Theta}_t\|_\infty)$ is non-decreasing with t. This hinders any conclusion on the convergence.

The discussion of this naive strategy is the starting point to understand what goes wrong in this case and to motivate our different approach. The main problem with a *global* comparison between the exact and the stochastic paths is that the iteration of triangle inequality bounding the discrepancy between Θ_t and $\tilde{\Theta}_t$ prodigally collects all the previous estimates, accumulating inaccuracy with time. What we need is a *local* approach. This can be achieved by performing a deterministic gradient descent step starting at the last point of the stochastic evolution, as depicted in Figure 21 on the right.

$$\begin{cases} \Theta_{t+1} = \Theta_t - \eta g^{(t)}(\Theta_t) & \text{stochastic gradient descent,} \\ \tilde{\Theta}_{t+1} = \Theta_t - \eta \nabla_\Theta \mathcal{L}(\Theta_t) & \text{locally deterministic gradient descent.} \end{cases} \tag{5.16}$$

Now, the variance of $g^{(t)}(\Theta_t)$ must be chosen so that $\|\Theta_{t+1} - \tilde{\Theta}_{t+1}\|_\infty$ is small enough to fulfill these criteria:

1. $\mathcal{L}(\Theta_{t+1})$ is sufficiently close to $\mathcal{L}(\tilde{\Theta}_{t+1})$ so that a decrease with respect to the previous cost $\mathcal{L}(\Theta_t)$ is ensured;
2. the distance between Θ_{t+1} and $\tilde{\Theta}_{t+1}$ is a summable function of t which goes to zero as the number of qubits increases, so that the lazy training result is still valid.

This will be the strategy to prove Theorem 5.4, which is the most important result of the article together with Theorem 5.9.

Theorem 5.4 (Trainability of the model) *Let us suppose that Assumption 4.3 holds. Let us furthermore assume that*

$$\lim_{m\to\infty} \frac{Lm|\mathcal{M}|^4|\mathcal{N}|^2}{N^3(m)} = 0. \tag{5.17}$$

We suppose to train our model according to the gradient descent described in Assumption 5.3 for a training time $T \in \mathbb{N}$ with learning rate $\eta = \frac{n}{N_K(m)}\eta_0$, where

$$\eta_0 < \frac{2}{\lambda^K_{\min} + \lambda^K_{\max}}. \tag{5.18}$$

We furtherly assume the following condition on the variance of the estimator of the gradient: for any choice of $t \leq T$ and $\xi_k \in \mathbb{R}$ and $\Theta_k \in \mathbb{R}^{|\Theta|}$,

$$\begin{aligned}\mathrm{Var}\Big[g_i^{(t)}(\Theta) \;\Big|\; g^{(k)}(\Theta_k) = \xi_k \quad \forall\, k < t\Big] \\ \leq c_0 \eta_0^2 \frac{\left(\lambda_{\min}^K\right)^4}{n^2} \frac{N_K^2(m) N^2(m)}{|\mathcal{M}|^2 |\Theta|^3} \frac{\delta}{(t+1)^2} \mathcal{L}(\Theta),\end{aligned} \tag{5.19}$$

where $c_0 = \frac{1}{864\pi^2}$.
For any $\delta > 0$, there exist $\bar{m} \in \mathbb{N}$ and some constants $R_0(\delta)$, $R_1(\delta)$, $R_2(\delta)$ such that for all $m \geq \bar{m}$ the following inequalities hold with probability at least $1 - \delta$ over random initialization and stochastic training:

$$\mathcal{L}(\Theta_t) \leq \frac{R_0^2}{2} \log(2n) \left(1 - \frac{\eta_0 \lambda_{\min}^K}{3}\right)^{2t}, \tag{5.20}$$

$$\begin{aligned}\|\Theta_t - \Theta_0\|_\infty \leq R_1 \eta_0 \lambda_{\min}^K \sqrt{\log(2n)} \frac{N(m)}{|\mathcal{M}||\Theta|} \\ + \frac{R_2}{\lambda_{\min}^K} n \sqrt{\log(2n)} \frac{|\mathcal{M}|}{N_K(m) N(m)},\end{aligned} \tag{5.21}$$

$$\begin{aligned}\sup_{x \in \mathcal{X}} |f(\Theta_t, x) - f^{\mathrm{lin}}(\Theta_t, x)| \leq 2 R_1^2 \eta_0^2 \left(\lambda_{\min}^K\right)^2 \log(2n) \frac{|\mathcal{N}| N(m)}{Lm} \\ + \frac{2R_2^2}{\left(\lambda_{\min}^K\right)^2} n^2 \log(2n) \frac{Lm |\mathcal{M}|^4 |\mathcal{N}|}{N_K^2(m) N^3(m)},\end{aligned} \tag{5.22}$$

for any $t \leq T$.

Proof See subsection 5.5. □

Remark 5.3 (Dependence of the variance on $\mathcal{L}(\Theta)$ and on t) As already introduced in subsection 5.1, the variance of the estimator of the gradient must be sufficiently small so that the exact value of the gradient is not completely blurred by the statistical noise. In the proof of Theorem 5.4 we show that a small cost function implies a small gradient. So, the cost function in the bound (5.19) ensures that the variance is exponentially small (in time) when the cost and its gradient are exponentially small. Furthermore, an explicit dependence on t appears in (5.19). This is due to the iterative comparison of the stochastic gradient descent with the exact gradient descent of the strategy we discussed in the introduction of this subsection. Indeed, we need the probability that the estimated gradient is too far from the exact gradient to be increasingly small along the minimization procedure. Otherwise, we could not provide a union bound ensuring that the sum of the probabilities that estimated gradient is not close enough to the exact gradient is small (see (5.113)).

In the following proposition we explain why the bound on the variance (5.19) implies that a polynomial number of measurements (with respect to the number of the qubits) is sufficient to

Proposition 5.5 (A polynomial number of measures) *For any $\epsilon, \delta > 0$, let $M(\epsilon, \delta)$ be the number of measurements required by the gradient-descent algorithm of Theorem 5.4 to get a cost less than ϵ with probability at least $1 - \delta$. Then,*

$$M(\epsilon, \delta) \leq C(\delta, \epsilon, \lambda_{\min}^K, n) \left(\frac{L\, m\, |\mathcal{M}|^4}{N^3(m)}\right)^{\frac{4}{3}} L^{\frac{8}{3}} m^{\frac{14}{3}}, \tag{5.23}$$

where the term in parentheses tends to 0 for $m \to \infty$ from the hypothesis (5.17) of Theorem 5.4, and the remaining terms grow at most polynomially in m whenever the number of layers L grows polynomially with m.

Proof We recall that, if X and Y are independent random variables, then

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y], \tag{5.24}$$

therefore, if $X_1, \ldots, X_M$ are i.i.d. random variables, then

$$\mathrm{Var}\left[\frac{1}{M} \sum_{k=1}^{M} X_k\right] = \frac{\mathrm{Var}[X_1]}{M}. \tag{5.25}$$

By Assumption 2.2, the generators of the unitaries encoding the trainable parameters have spectrum $\{-1,+1\}$. Then, there is a simple strategy to compute the derivative $\partial_{\theta_i} f(\Theta,x)$. In quantum circuits like these, indeed, the derivatives of the expectation value of any observable with respect to any parameter can be written in terms of other expectation values computed for "shifted" parameters. These identities were introduced in [62] (see in particular (14)) and they are known as parameter shift rules:

$$\frac{\partial}{\partial\theta_i} f(\Theta,x) = f(\Theta+\Delta^{(i)},x) - f(\Theta-\Delta^{(i)},x) \qquad \Delta_j^{(i)} := \begin{cases} \frac{\pi}{4} & j=i \\ 0 & j\neq i \end{cases} \tag{5.26}$$

so that an unbiased estimator $h_i^{(t,1)}(\Theta,x)$ of $\partial_{\theta_i} f(\Theta,x)$ can be constructed by measuring the output of the circuit at 2 different points

$$h_i^{(t,1)}(\Theta,x) = h_i^{(t,1,+)}(\Theta,x) - h_i^{(t,1,-)}(\Theta,x), \tag{5.27}$$

where $h_i^{(t,1,\pm)}(\Theta,x)$ are the outputs of the *mutually independent* measurements of $\mathcal{O}$ on $\left|\psi(\Theta\pm\Delta^{(i)},x)\right\rangle = U(\Theta\pm\Delta^{(i)},x)\left|0^m\right\rangle$:

$$\begin{aligned} \mathbb{E}\left[h_i^{(t,1)}(\Theta,x)\right] &= \mathbb{E}\left[h_i^{(t,1,+)}(\Theta,x) - h_i^{(t,1,-)}(\Theta,x)\right] \\ &= h_i^{(t,1,+)}(\Theta,x) - h_i^{(t,1,-)}(\Theta,x) = \frac{\partial}{\partial\theta_i} f(\Theta,x). \end{aligned} \tag{5.28}$$

All the expectation values and all the variances are meant to be conditioned on any possible story of the previous estimators. Iterating this procedure $M^{(t)}$ times, we have a family of unbiased and independent estimators $\{h_i^{(t,k)}(\Theta,x)\}_{k\in\{1,\dots,M^{(t)}\}}$ of $\frac{\partial}{\partial\theta_i} f(\Theta,x)$. Since, by Assumption 2.2, $\mathcal{O}$ is a sum of m observables with spectrum in $[-1,1]$, a very rough uniform bound on the variance of $h_i^{(t,1,j)}(\Theta+\Delta^{(j)},x)$ can be easily provided:

$$\mathrm{Var}\left[h_i^{(t,k,j)}(\Theta+\Delta^{(j)},x)\right] \leq \frac{m^2}{N^2(m)}, \tag{5.29}$$

whence

$$\mathrm{Var}\left[h_i^{(t,k)}(\Theta,x)\right] \leq \frac{2m^2}{N^2(m)}. \tag{5.30}$$

Therefore,

$$g_i^{(t)}(\Theta,x) = \frac{1}{M^{(t)}} \sum_{k=1}^{M^{(t)}} h_i^{(t,k)}(\Theta,x) \tag{5.31}$$

is an unbiased estimator of $\frac{\partial}{\partial\theta_i} f(\Theta,x)$ with variance

$$\mathrm{Var}\left[g_i^{(t)}(\Theta,x)\right] \leq \frac{1}{M^{(t)}} \frac{2m^2}{N^2(m)}. \tag{5.32}$$

By (5.19) we notice that, in the physical implementation of the quantum neural network, the variance required at each time step is of the form

$$\mathrm{Var}\Big[g_i^{(t)}(\Theta_t) \;\Big|\; g^{(k)}(\Theta_k)=\xi_k \quad \forall\, k<t\Big] \leq c(\delta,\lambda_{\min}^K,n)\frac{N_K^2(m)N^2(m)}{|\mathcal{M}|^2L^3m^3}\frac{\mathcal{L}(\Theta_t)}{(t+1)^2}. \tag{5.33}$$

This is ensured if the number of measurements $M^{(t)}$ is sufficiently large

$$\frac{1}{M^{(t)}}\frac{2m^2}{N^2(m)} \leq c(\delta,\lambda_{\min}^K,n)\frac{N_K^2(m)N^2(m)}{|\mathcal{M}|^2L^3m^3}\frac{\mathcal{L}(\Theta_t)}{(t+1)^2}, \tag{5.34}$$

i.e.

$$M^{(t)} = \frac{2}{c(\delta,\lambda_{\min}^K,n)}\frac{|\mathcal{M}|^2L^3m^5}{N_K^2(m)N^4(m)}\frac{(t+1)^2}{\mathcal{L}(\Theta_t)} \tag{5.35}$$

is sufficient. Since this procedure must be repeated for each parameter, the total number of measurements required at time t is

$$M_{tot}^{(t)} = |\Theta|M^{(t)} = LmM^{(t)}. \tag{5.36}$$

To get ϵ-close to the minimum of the cost, by (5.20) the number of time steps required is, at most,

$$T_{\max}^{(\epsilon)} = \frac{\log\left(\frac{1}{\epsilon}\right)+\log\left(R_0^2\log(2n)\right)}{-2\log\left(1-\frac{\eta_0\lambda_{\min}^K}{3}\right)}. \tag{5.37}$$

So the total number of measurements required is bounded by

$$M(\epsilon,\delta) \leq \sum_{t=1}^{T_{\max}^{(\epsilon)}} M_{tot}^{(t)} \leq \frac{2T_{\max}^{(\epsilon)}(T_{\max}^{(\epsilon)}+1)^2}{c(\delta,\lambda_{\min}^K,n)\epsilon}\frac{|\mathcal{M}|^2L^4m^6}{N_K^2(m)N^4(m)}. \tag{5.38}$$

Setting

$$C(\delta,\epsilon,\lambda^K_{\min},n) \coloneqq \frac{2T^{(\epsilon)}_{\max}(T^{(\epsilon)}_{\max}+1)^2}{c(\delta,\lambda^K_{\min},n)\epsilon} \tag{5.39}$$

we get

$$\begin{aligned} M(\epsilon,\delta) &\leq C(\delta,\epsilon,\lambda^K_{\min},n)\,\frac{|\mathcal{M}|^2L^4m^6}{N_K^2(m)N^4(m)} \overset{(a)}{\leq} C(\delta,\epsilon,\lambda^K_{\min},n)\left(\frac{L\,m\,|\mathcal{M}|^4}{N^3(m)}\right)^{\frac{4}{3}}\frac{L^{\frac{8}{3}}m^{\frac{14}{3}}}{|\mathcal{M}|^{\frac{10}{3}}} \\ &\leq C(\delta,\epsilon,\lambda^K_{\min},n)\left(\frac{L\,m\,|\mathcal{M}|^4}{N^3(m)}\right)^{\frac{4}{3}}L^{\frac{8}{3}}m^{\frac{14}{3}}\,, \end{aligned} \tag{5.40}$$

where (a) follows from Lemma 4.5. The claim follows. □

Theorem 5.4 is the fundamental result to prove that also in the noisy case the model is a Gaussian process during the training in the limit of many qubits. However, we still need to verify that the noisy trajectory in the parameter space is close to the exact trajectory with the linearized evolution equation. This result is ensured by the following statements, which require a further suppression $\xi(m)\to 0$ in the bound for variance of the estimator for the gradient of the cost function; we are going to choose $\xi(m)$ in order to ensure the rate of convergence needed to obtain the convergence to a Gaussian process. We will also replace δ with $\delta/4$: this is just a technical detail which will be used in the proof for a union bound.

Lemma 5.6 (The parameters differ at the second order) *Let us assume a condition on the variance of the estimator of the gradient stronger than (5.19) by a factor $\xi(m)/4$: for any choice of $t\leq T$ and $\xi_k\in\mathbb{R}$ and $\Theta_k\in\mathbb{R}^{|\Theta|}$,*

$$\begin{aligned} &\mathrm{Var}\Big[g^{(t)}_i(\Theta)\;\Big|\;g^{(k)}(\Theta_k)=\xi_k\quad\forall\,k<t\Big] \\ &\qquad\leq c_0\eta_0^2\,\frac{\left(\lambda^K_{\min}\right)^4}{n^2}\,\frac{N_K^2(m)N^2(m)}{|\mathcal{M}|^2|\Theta|^3}\xi(m)\,\frac{\delta/4}{(t+1)^2}\,\mathcal{L}(\Theta), \end{aligned} \tag{5.41}$$

where $c_0=\frac{1}{864\pi^2}$. Then, for any fixed $\delta>0$, there exist some constant C_1,C_2 and an integer $\bar m\in\mathbb{N}$ such that, for any $m\geq\bar m$,

$$\left\|\Theta_t-\Theta^{\mathrm{lin}}_t\right\|_\infty \leq \left(\frac{C_1}{(\lambda^K_{\min})^3}+C_2\right)n^3\log(2n)\frac{Lm|\mathcal{M}|^5|\mathcal{N}|^3}{N^4(m)}\log N(m)\frac{1}{2}\left(1+N(m)\sqrt{\xi(m)}\right) \tag{5.42}$$

with probability at least $1-\delta$.

Proof See subsection 5.6. □

Now it is clear why we need the further suppression $\xi(m)$ in the bound for the variance of the estimator of the gradient of the cost function, as we will discuss more in detail in Corollary (5.7) and in Theorem (5.8): the divergent term $N(m)$ must be cancelled by $\xi(m)$ in order to ensure the following statements.

Corollary 5.7 *If we choose, in particular*

$$\xi(m)=\frac{1}{N^2(m)}, \tag{5.43}$$

which means requiring that

$$\begin{aligned} &\mathrm{Var}\Big[g^{(t)}_i(\Theta)\;\Big|\;g^{(k)}(\Theta_k)=\xi_k\quad\forall\,k<t\Big] \\ &\qquad\leq c_0\eta_0^2\,\frac{\left(\lambda^K_{\min}\right)^4}{n^2}\,\frac{N_K^2(m)}{|\mathcal{M}|^2|\Theta|^3}\,\frac{\delta/4}{(t+1)^2}\,\mathcal{L}(\Theta), \end{aligned} \tag{5.44}$$

where $c_0=\frac{1}{864\pi^2}$, then, for any $\delta>0$, there exist C_1,C_2 and m_0 such that, for any $m\geq m_0$,

$$\left\|\Theta_t-\Theta^{\mathrm{lin}}_t\right\|_\infty \leq \left(\frac{C_1}{(\lambda^K_{\min})^3}+C_2\right)n^3\log(2n)\frac{Lm|\mathcal{M}|^5|\mathcal{N}|^3}{N^4(m)}\log N(m) \tag{5.45}$$

with probability at least $1-\delta$.

Now we have all the ingredients to state the generalization of Theorem 4.15 to the discrete time setting with statistical noise.

Theorem 5.8 (The original evolution is close to the linear evolution) *Let us suppose that the hypotheses of the previous lemmas are satisfied; in particular, we ask (5.44) for the variance of the estimator of the gradient of the cost function. Then, in the presence or in the absence of statistical noise in the gradient descent, we have the following bound: for any $\delta > 0$ there exist some constants C_3 and C_4 and a number of qubits $m_0 \in N$ such that, for any $m \geq m_0$*

$$|f(\Theta_t, x) - f^{\text{lin}}(\Theta_t^{\text{lin}}, x)| \leq \left(\frac{C_3}{(\lambda_{\min}^K)^3} + C_4 \right) n^3 \log(2n) \frac{L^2 m^2 |\mathcal{M}|^6 |\mathcal{N}|^4}{N^5(m)} \log N(m) \tag{5.46}$$

with probability at least $1 - \delta$.

Proof See subsection 5.7. □

Theorem 5.9 (Quantum neural networks as Gaussian processes – noisy setting) *Let us assume that a circuit satisfying Assumption 4.3 and such that*

$$\lim_{m \to \infty} \frac{L^2 m^2 |\mathcal{M}|^6 |\mathcal{N}|^4}{N^5(m)} \log N(m) = 0 \tag{5.47}$$

is randomly initialized according to Assumption 3.6 and is trained using the noisy gradient descent with the following bound on the variance of the estimator of the gradient of the cost function: for any choice of $t \leq T$ and $\xi_k \in \mathbb{R}$ and $\Theta_k \in \mathbb{R}^{|\Theta|}$,

$$\begin{aligned} &\text{Var}\left[g_i^{(t)}(\Theta) \;\middle|\; g^{(k)}(\Theta_k) = \xi_k \quad \forall\, k < t \right] \\ &\qquad \leq c_0 \eta_0^2 \frac{\left(\lambda_{\min}^K\right)^4}{n^2} \frac{N_K^2(m)}{|\mathcal{M}|^2 |\Theta|^3} \frac{\delta/4}{(t+1)^2} \mathcal{L}(\Theta). \end{aligned} \tag{5.48}$$

Then, in the limit of infinitely many qubits $m \to \infty$, $\{f(\Theta_t, x)\}_{x \in \mathcal{X}}$ converges in distribution to a Gaussian process $\{f_t^{(\infty)}(x)\}_{x \in \mathcal{X}}$ with mean and covariance

$$\mu_t(x) = \bar{K}(x, X^T) \bar{K}^{-1} \left(\mathbb{1} - \left(\mathbb{1} - \eta_0 \bar{K} \right)^t \right) Y \tag{5.49}$$

$$\begin{aligned} \mathcal{K}_t(x, x') = \mathcal{K}_0(x, x'), \\ &- \bar{K}(x, X^T) \bar{K}^{-1} \left(\mathbb{1} - \left(\mathbb{1} - \eta_0 \bar{K} \right)^t \right) \mathcal{K}_0(X, x') \\ &- \bar{K}(x', X^T) \bar{K}^{-1} \left(\mathbb{1} - \left(\mathbb{1} - \eta_0 \bar{K} \right)^t \right) \mathcal{K}_0(X, x) \\ &+ \bar{K}(x, X^T) \bar{K}^{-1} \left(\mathbb{1} - \left(\mathbb{1} - \eta_0 \bar{K} \right)^t \right) \times \\ &\qquad \times \mathcal{K}_0(X, X^T) \left(\mathbb{1} - \left(\mathbb{1} - \eta_0 \bar{K} \right)^t \right) \bar{K}^{-1} \bar{K}(X, x'). \end{aligned} \tag{5.50}$$

Proof See subsection 5.8. □

5.4 Proof of Theorem 5.1

Lemma 5.10 (Lagrange theorem for vector valued functions in several variables) *Let $\mathcal{U} \subseteq \mathbb{R}^m$ be a compact convex set and let $f : \mathcal{U} \to \mathbb{R}^n$ be continuously differentiable. Then*

$$f(y) - f(x) = \left[\int_0^1 df(x + t(y - x)) dt \right] (y - x). \tag{5.51}$$

Proof See e.g. [63]. □

The following lemma is crucial both for the deterministic and the stochastic case. We show that, if we are asymptotically close enough to the parameters at initialization, then a gradient descent step contracts the distance between the output of the model and the labels provided by the dataset.

Let $B_r(\Theta_0) = \{\Theta : \|\Theta - \Theta_0\|_\infty < r\}$ be the ball of center Θ_0 and radius r.

Lemma 5.11 (Single step of gradient descent) *Let us assume that the hypothesis of Theorem 4.7 is satisfied and that Assumption 4.3 holds. Let $\delta > 0$, $\zeta \in (0,1)$ and suppose*

$$\eta = \frac{n}{N_K(m)}\eta_0, \qquad \eta_0 < \frac{2}{\lambda^K_{\min} + \lambda^K_{\max}}. \tag{5.52}$$

Furthermore, let $r(m) : \mathbb{N} \to \mathbb{R}^+$ such that

$$\lim_{m\to\infty} \frac{\Sigma_1 |\mathcal{M}|^2 |\mathcal{N}|}{N_K(m) N^2(m)} r(m) = 0. \tag{5.53}$$

There exists $\bar{m} \in \mathbb{N}$ such that, with probability at least $1-\delta$ over random initialization, given any Θ and Θ' both in $B_{r(m)}(\Theta_0)$ and satisfying

$$\Theta' - \Theta = -\eta \nabla_\Theta \mathcal{L}(\Theta), \tag{5.54}$$

where $\mathcal{L}$ is the MSE loss, the following inequality holds for any $m \geq \bar{m}$:

$$\|f(\Theta', X) - Y\|_2 \leq \left(1 - \zeta\eta_0 \lambda^K_{\min}\right) \|f(\Theta, X) - Y\|_2. \tag{5.55}$$

Proof Let us start rewriting

$$\|f(\Theta', x) - Y\|_2 = \|(f(\Theta', x) - f(\Theta, x)) + (f(\Theta, x) - Y)\|_2. \tag{5.56}$$

Calling[16]

$$DF(\tau) \equiv \begin{pmatrix} \nabla_\Theta f(\Theta + \tau(\Theta' - \Theta), x^{(1)})^T \\ \nabla_\Theta f(\Theta + \tau(\Theta' - \Theta), x^{(2)})^T \\ \vdots \\ \nabla_\Theta f(\Theta + \tau(\Theta' - \Theta), x^{(n)})^T \end{pmatrix} \tag{5.57}$$

we can apply Lagrange theorem 5.10 to obtain

$$\begin{aligned} \|f(\Theta', X) - Y\|_2 &= \left\| \left[\int_0^1 DF(\tau) d\tau \right] (\Theta' - \Theta) + (f(\Theta, X)) - Y) \right\|_2 \\ &= \left\| \int_0^1 \big(DF(\tau)(\Theta' - \Theta) + (f(\Theta, X) - Y) \big) d\tau \right\|_2 \\ &\leq \int_0^1 d\tau \left\| DF(\tau)(\Theta' - \Theta) + f(\Theta, X) - Y \right\|_2 \\ &= \left\| DF(\xi)(\Theta' - \Theta) + f(\Theta, X) - Y \right\|_2, \end{aligned} \tag{5.58}$$

where $\xi \in [0,1]$. By the definition of gradient descent (5.54), we also have

$$\begin{aligned} \|f(\Theta', X) - Y\|_2 &\leq \| - \frac{\eta_0}{N_K(m)} DF(\xi) \nabla_\Theta \mathcal{L}(\Theta) + f(\Theta, X) - Y \|_2 \\ &\leq \| - \frac{\eta_0}{N_K(m)} DF(\xi) \nabla_\Theta f(\Theta, X^T)(f(\Theta, X) - Y) + f(\Theta, X) - Y \|_2 \\ &= \left\| \left(\mathbb{1} - \frac{\eta_0}{N_K(m)} DF(\xi) \nabla_\Theta f(\Theta, X^T) \right) (f(\Theta, X) - Y) \right\|_2 \\ &\leq \left\| \left(\mathbb{1} - \frac{\eta_0}{N_K(m)} DF(\xi) \nabla_\Theta f(\Theta, X^T) \right) \right\|_{\mathcal{L}} \|f(\Theta, X) - Y\|_2. \end{aligned} \tag{5.59}$$

Calling $\tilde{\Theta} = \Theta + \xi(\Theta' - \Theta)$, by convexity of the ball,

$$\tilde{\Theta} \in B_{r(m)}(\Theta_0) \quad \text{because} \quad \begin{cases} \|\Theta - \Theta_0\|_\infty \leq r(m) \\ \|\Theta' - \Theta_0\|_\infty \leq r(m) \end{cases}. \tag{5.60}$$

[16] We use the convention of considering ∇f^T as a row vector.

Now we have to estimate

$$\begin{aligned}
&\left\| \left(\mathbb{1} - \frac{\eta_0}{N_K(m)} DF(\xi)(\nabla_\Theta f(\Theta, X^T) \right) \right\|_{\mathcal{L}} \\
&\quad = \left\| \left(\mathbb{1} - \frac{\eta_0}{N_K(m)} \sum_{i=1}^{Lm} \partial_{\theta_i} f(\tilde{\Theta}, X) \partial_{\theta_i} f(\Theta, X^T) \right) \right\|_{\mathcal{L}} \\
&\quad = \Bigg\| \Bigg(\mathbb{1} - \eta_0 \Big(K(X, X^T) - K(X, X^T) + \hat{K}_{\Theta_0}(X, X^T) \\
&\qquad\qquad - \hat{K}_{\Theta_0}(X, X^T) + \frac{1}{N_K(m)} \sum_{i=1}^{Lm} \partial_{\theta_i} f(\tilde{\Theta}, X) \partial_{\theta_i} f(\Theta, X^T) \Big) \Bigg) \Bigg\|_{\mathcal{L}} \\
&\quad \leq \| \mathbb{1} - \eta_0 K(X, X^T) \|_{\mathcal{L}} + \eta_0 \| K(X, X^T) - \hat{K}_{\Theta_0}(X, X^T) \|_{\mathcal{L}} \\
&\qquad + \eta_0 \left\| \sum_{i=1}^{Lm} \partial_{\theta_i} f(\Theta_0, X) \partial_{\theta_i} f(\Theta_0, X^T) - \sum_{i=1}^{Lm} \partial_{\theta_i} f(\tilde{\Theta}, X) \partial_{\theta_i} f(\Theta, X^T) \right\|_{\mathcal{L}} \\
&\quad \leq \| \mathbb{1} - \eta_0 K(X, X^T) \|_{\mathcal{L}} + \eta_0 \| K(X, X^T) - \hat{K}_{\Theta_0}(X, X^T) \|_{\mathcal{L}} \\
&\qquad + \frac{\eta_0}{N_K(m)} \left\| \sum_{i=1}^{Lm} \left(\partial_{\theta_i} f(\Theta_0, X) - \partial_{\theta_i} f(\tilde{\Theta}, X) \right) \partial_{\theta_i} f(\Theta_0, X^T) \right\|_{\mathcal{L}} \\
&\qquad + \frac{\eta_0}{N_K(m)} \left\| \sum_{i=1}^{Lm} \partial_{\theta_i} f(\tilde{\Theta}, X) \left(\partial_{\theta_i} f(\Theta_0, X^T) - \partial_{\theta_i} f(\Theta, X^T) \right) \right\|_{\mathcal{L}}.
\end{aligned} \tag{5.61}$$

Each term can be bounded as follows. Let us start with

$$\| \mathbb{1} - \eta_0 K(X, X^T) \|_{\mathcal{L}} \leq \max_{\lambda \in \mathrm{Spec} K} |1 - \eta_0 \lambda|. \tag{5.62}$$

Since $f(\lambda) = |1 - \eta_0 \lambda|$ is convex,

$$\| \mathbb{1} - \eta_0 K(X, X^T) \|_{\mathcal{L}} \leq \max\{ |1 - \eta_0 \lambda^K_{\min}|, |1 - \eta_0 \lambda^K_{\max}| \} \tag{5.63}$$

and, since $\eta_0 < \frac{2}{\lambda^K_{\min} + \lambda^K_{\max}}$,

$$1 - \eta_0 \lambda^* = 0 \quad \Longleftrightarrow \quad \lambda^* = \frac{1}{\eta_0} > \frac{\lambda^K_{\min} + \lambda^K_{\max}}{2}. \tag{5.64}$$

Therefore, we can conclude that

$$\| \mathbb{1} - \eta_0 K(X, X^T) \|_{\mathcal{L}} \leq 1 - \eta_0 \lambda^K_{\min} \tag{5.65}$$

as illustrated in Figure 22. Now we consider

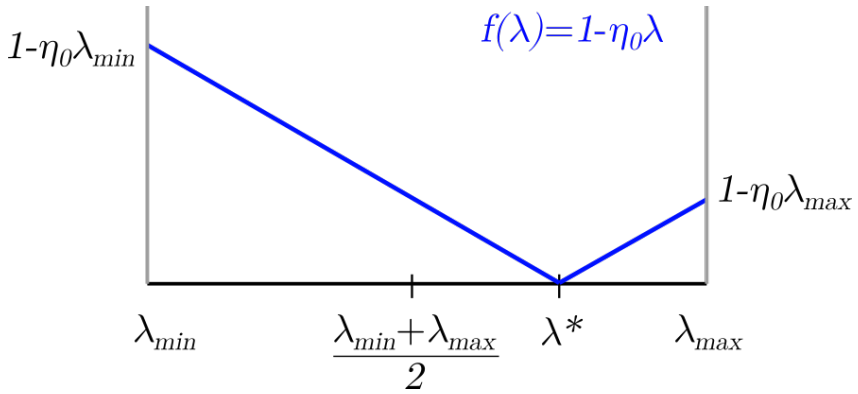


Fig. 22 Behaviour of $f(\lambda)$

$$\eta_0 \| K(X, X^T) - \hat{K}_{\Theta_0}(X, X^T) \|_{\mathcal{L}} \leq \| K(X, X^T) - \hat{K}_{\Theta_0}(X, X^T) \|_F. \tag{5.66}$$

We can claim, by Theorem 4.7 and Lemma 4.24, the existence of $m_1 \in \mathbb{N}$ such that

$$\| K(X, X^T) - \hat{K}_{\Theta_0}(X, X^T) \|_F \leq \left(\frac{1 - \zeta}{2} \right) \lambda^K_{\min} \qquad \forall\, m \geq m_1 \tag{5.67}$$

with probability at least $1-\delta$.
Finally, using Lemma 4.20,

$$\begin{aligned}
&\left\| \sum_{i=1}^{Lm} \left(\partial_{\theta_i} f(\Theta_0, X) - \partial_{\theta_i} f(\tilde{\Theta}, X)\partial_{\theta_i} f(\Theta_0, X^T)\right)\right\|_{\mathcal{L}} \\
&\quad \leq \left\| \sum_{i=1}^{Lm} \left(\partial_{\theta_i} f(\Theta_0, X) - \partial_{\theta_i} f(\tilde{\Theta}, X)\partial_{\theta_i} f(\Theta_0, X^T)\right)\right\|_{F} \\
&\quad \leq \sqrt{\sum_{j,j'=1}^{n} \left(\sum_{i=1}^{Lm} \left(\partial_{\theta_i} f(\Theta_0, x^{(j)}) - \partial_{\theta_i} f(\tilde{\Theta}, x^{(j)})\partial_{\theta_i} f(\Theta_0, x^{(j')})\right)\right)^2} \\
&\quad \leq \sqrt{\sum_{j,j'=1}^{n} \left(\sup_{i'} \left|\partial_{\theta_{i'}} f(\Theta_0, x^{(j)}) - \partial_{\theta_{i'}} f(\tilde{\Theta}, x^{(j)})\right| \sum_{i=1}^{Lm} \left|\partial_{\theta_i} f(\Theta_0, x^{(j')})\right|\right)^2} \\
&\quad \leq \sup_x \sup_{i'} \left|\partial_{\theta_{i'}} f(\Theta_0, x) - \partial_{\theta_{i'}} f(\tilde{\Theta}, x)\right| \sqrt{\sum_{j,j'=1}^{n} \left(\sum_{i=1}^{Lm} \left|\partial_{\theta_i} f(\Theta_0, x^{(j')})\right|\right)^2} \\
&\quad \leq \sup_x \sup_{i'} \left|\partial_{\theta_{i'}} f(\Theta_0, x) - \partial_{\theta_{i'}} f(\tilde{\Theta}, x)\right| \sup_{x'} \sum_{i=1}^{Lm} \left|\partial_{\theta_i} f(\Theta_0, x'))\right| \sqrt{\sum_{j,j'=1}^{n} 1} \\
&\quad \leq n \sup_x \left\|\nabla_\Theta f(\Theta_0, x) - \nabla_\Theta f(\tilde{\Theta}, x)\right\|_\infty \sup_{x'} \left\|\nabla_\Theta f(\Theta_0, x')\right\|_1 \\
&\quad \leq n \cdot 4 \frac{|\mathcal{M}|^2|\mathcal{N}|}{N(m)} \|\Theta_0 - \tilde{\Theta}\|_\infty \cdot \frac{2\Sigma_1}{N(m)} \leq 8n \frac{\Sigma_1 |\mathcal{M}|^2|\mathcal{N}|}{N^2(m)} r(m).
\end{aligned} \tag{5.68}$$

Hence

$$\begin{aligned}
&\left\| \left(\mathbb{1} - \frac{\eta_0}{N_K(m)} DF(\xi)(\nabla_\Theta f(\Theta, X^T)\right)\right\|_{\mathcal{L}} \\
&\quad \leq 1 - \eta_0 \lambda_{\min}^K + \left(\frac{1-\zeta}{2}\right) \eta_0 \lambda_{\min}^K + 8n\eta_0 \frac{\Sigma_1 |\mathcal{M}|^2|\mathcal{N}|}{N_K(m)N^2(m)} r(m).
\end{aligned} \tag{5.69}$$

Because of (5.53), there exist $m_2 \in \mathbb{N}$ such that

$$8n\eta_0 \frac{\Sigma_1 |\mathcal{M}|^2|\mathcal{N}|}{N_K(m)N^2(m)} r(m) \leq \left(\frac{1-\zeta}{2}\right) \eta_0 \lambda_{\min}^K \qquad \forall m \geq m_2. \tag{5.70}$$

This yields

$$\begin{aligned}
&\left\| \left(\mathbb{1} - \frac{\eta_0}{N_K(m)} DF(\xi)(\nabla_\Theta f(\Theta, X^T)\right)\right\|_{\mathcal{L}} \\
&\quad \leq 1 - \eta_0 \lambda_{\min}^K + (1-\zeta)\eta_0 \lambda_{\min}^K = 1 - \zeta\eta_0 \lambda_{\min}^K.
\end{aligned} \tag{5.71}$$

Therefore

$$\|f(\Theta', X) - Y\|_2 \leq (1 - \zeta\eta_0 \lambda_{\min}^K) \|f(\Theta, X) - Y\|_2 \tag{5.72}$$

for all $m \geq \bar{m} = \max\{m_1, m_2\}$.

□

Now we have all the ingredients to prove Theorem 5.1.
Because of Corollary 4.16, we can ask

$$\mathbb{P}\left(\|F(0) - Y\|_2 < R\right) \geq 1 - \frac{\delta}{2} \quad \forall m \geq m_0 \text{ for some } R = \sqrt{n\log(2n)} R_0 \tag{5.73}$$

with probability at least $1 - \frac{\delta}{2}$.
As in the proof of the Theorem 4.10, we introduce

$$\rho(m) = \frac{6R_0}{\lambda_{\min}^K} n\sqrt{\log(2n)} \frac{|\mathcal{M}|}{N_K(m)N(m)}. \tag{5.74}$$

We prove (5.8) and (5.9) by induction. If $t = 0$, then

$$\|F(0) - Y\|_2 < R \tag{5.75}$$

holds because of the previous assumption; (5.9) does not have to be proved for $t = 0$. Let us suppose that (5.8) and (5.9) hold for any $t \leq t^*$. Recalling that

$$\Theta_{t+1} - \Theta_t = -\frac{\eta_0}{N_K(m)} \nabla_\Theta f(\Theta_t, X) \cdot (F(t) - Y) \tag{5.76}$$

and using Lemma 4.20, we have

$$\begin{aligned} |\theta_i(t+1) - \theta_i(t)| &\leq \left| \frac{\eta_0}{N_K(m)} \partial_{\theta_i} f(\Theta_t, X) \cdot (F(t) - Y) \right| \\ &\leq \frac{\eta_0}{N_K(m)} 2\sqrt{n} \frac{|\mathcal{M}|}{N(m)} \|F(t) - Y\|_2 \end{aligned} \tag{5.77}$$

Hence, because (5.8) holds for any $t \leq t^*$,

$$\begin{aligned} |\theta_i(t+1) - \theta_i(t)| &\leq \frac{\eta_0}{N_K(m)} 2\sqrt{n} \frac{|\mathcal{M}|}{N(m)} \sqrt{n \log(2n)} R_0 \left(1 - \frac{1}{3}\eta_0 \lambda_{\min}^K\right)^t \\ &= 2R_0 \eta_0 n \sqrt{\log(2n)} \frac{|\mathcal{M}|}{N_K(m) N(m)} \left(1 - \frac{1}{3}\eta_0 \lambda_{\min}^K\right)^t \qquad \forall\, t \leq t^*. \end{aligned} \tag{5.78}$$

Therefore, using the triangle inequality and considering the supremum over i of the previous inequalities,

$$\begin{aligned} \|\Theta_{t^*+1} - \Theta_0\|_\infty &\leq \sum_{k=0}^{t^*} \|\Theta_{k+1} - \Theta_k\|_\infty \\ &\leq 2R_0 n \sqrt{\log(2n)} \eta_0 \frac{|\mathcal{M}|}{N_K(m) N(m)} \sum_{k=0}^{t^*} \left(1 - \frac{1}{3}\eta_0 \lambda_{\min}^K\right)^k \\ &\leq 2R_0 n \sqrt{\log(2n)} \eta_0 \frac{|\mathcal{M}|}{N_K(m) N(m)} \sum_{k=0}^{\infty} \left(1 - \frac{1}{3}\eta_0 \lambda_{\min}^K\right)^k \\ &= \frac{6R_0}{\lambda_{\min}^K} n \sqrt{\log(2n)} \frac{|\mathcal{M}|}{N_K(m) N(m)} = \rho(m). \end{aligned} \tag{5.79}$$

This means that (5.9) holds also for $t = t^* + 1$. Now, let us notice that

$$\lim_{m \to \infty} \frac{\Sigma_1 |\mathcal{M}|^2 |\mathcal{N}|}{N_K(m) N^2(m)} \rho(m) = \lim_{m \to \infty} \frac{6R_0}{\lambda_{\min}^K} n \sqrt{\log(2n)} \frac{\Sigma_1 |\mathcal{M}|^3 |\mathcal{N}|}{N_K^2(m) N^3(m)} = 0 \tag{5.80}$$

by hypothesis. We invoke[17] Lemma 5.11 with probablity to fail $\frac{\delta}{2}$ instead of δ and choosing the following parameters:

$$r(m) = \rho(m), \qquad \Theta' = \Theta_{t^*+1}, \qquad \Theta = \Theta_{t^*}, \qquad \zeta = \frac{1}{3}. \tag{5.81}$$

Therefore

$$\|F(t^* + 1) - Y\|_2 \leq R_0 \sqrt{n \log(2n)} \left(1 - \frac{\eta_0 \lambda_{\min}^K}{3}\right)^{t^*+1} \tag{5.82}$$

which is (5.8) for $t = t^* + 1$.

[17] Since $r(m)$ and ζ are always the same in the proof by induction, the lemma has to be claimed once.

5.5 Proof of Theorem 5.4

Lemma 5.12 *The hypothesis*

$$\lim_{m\to\infty} \frac{Lm|\mathcal{M}|^4|\mathcal{N}|^2}{N^3(m)} = 0 \tag{5.83}$$

of Theorem 5.4 ensures that the hypotheses of Theorem 3.7, of Theorem 4.7 and of Theorem 5.1 are satisfied, i.e.

$$\lim_{m\to\infty} \frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3(m)} = 0, \tag{5.84}$$

$$\lim_{m\to\infty} \frac{1}{N_K^2(m)} \frac{\Sigma_2|\mathcal{M}|^2|\mathcal{N}|^2}{N^4(m)} = 0, \tag{5.85}$$

$$\lim_{m\to\infty} \frac{\Sigma_1|\mathcal{M}|^2|\mathcal{N}|}{N_K^2(m)N^3(m)} = 0. \tag{5.86}$$

Proof It is easy to see that

$$0 \le \lim_{m\to\infty} \frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3(m)} \le \lim_{m\to\infty} \frac{Lm|\mathcal{M}|^4|\mathcal{N}|^2}{N^3(m)} = 0, \tag{5.87}$$

$$0 \le \frac{1}{N_K^2(m)} \frac{\Sigma_2|\mathcal{M}|^2|\mathcal{N}|^2}{N^4(m)} \le \lim_{m\to\infty} \frac{Lm|\mathcal{M}|^4|\mathcal{N}|^2}{N^3(m)} = 0, \tag{5.88}$$

$$0 \le \lim_{m\to\infty} \frac{\Sigma_1|\mathcal{M}|^2|\mathcal{N}|}{N_K^2(m)N^3(m)} \le \lim_{m\to\infty} \frac{Lm|\mathcal{M}|^4|\mathcal{N}|^2}{N^3(m)} = 0. \tag{5.89}$$

□

Two simple lemmas are proved now, so that the computations will be simplified later.

Lemma 5.13 (An elementary inequality) *Let $x \in \mathbb{R}$ such that*

$$x^2 \le A + A' + Bx, \tag{5.90}$$

where $A, A', B > 0$. Then

$$x \le \sqrt{A} + \sqrt{A'} + B. \tag{5.91}$$

Proof The equality $x^2 = A + A' + Bx$ holds if and only if

$$x = x_1 = \frac{B + \sqrt{B^2 + 4(A + A')}}{2} \quad \text{or} \quad x = x_2 = \frac{B - \sqrt{B^2 + 4(A + A')}}{2}, \tag{5.92}$$

so the inequality $x^2 \le A + A' + Bx$ implies

$$x_2 \le x \le x_1. \tag{5.93}$$

Since

$$\begin{aligned} 2x_1 &= B + \sqrt{B^2 + 4(A + A')} = B + \sqrt{B^2 + (2\sqrt{A})^2 + (2\sqrt{A'})^2} \\ &\le B + B + 2\sqrt{A} + 2\sqrt{A'} = 2(B + \sqrt{A} + \sqrt{A'}), \end{aligned} \tag{5.94}$$

we have

$$x \le \sqrt{A} + \sqrt{A'} + B. \tag{5.95}$$

□

Lemma 5.14 (Another elementary inequality) *Let $x \in [0, 1]$. Then*

$$\frac{x}{4} \le \sqrt{1 - \frac{x}{2}} - \sqrt{1 - x}. \tag{5.96}$$

Proof First we notice that

$$\sqrt{1-\frac{x}{2}}-\sqrt{1-x}\geq 0. \tag{5.97}$$

Then we use that, for $x\in[0,1]$,

$$\sqrt{1-\frac{x}{2}}+\sqrt{1-x}\leq 2, \tag{5.98}$$

whence

$$\begin{aligned}\left(\sqrt{1-\frac{x}{2}}+\sqrt{1-x}\right)\left(\sqrt{1-\frac{x}{2}}-\sqrt{1-x}\right)&\leq 2\left(\sqrt{1-\frac{x}{2}}-\sqrt{1-x}\right)\\ \frac{1}{2}\left(1-\frac{x}{2}-(1-x)\right)&\leq\sqrt{1-\frac{x}{2}}-\sqrt{1-x}\\ \frac{x}{4}&\leq\sqrt{1-\frac{x}{2}}-\sqrt{1-x}.\end{aligned} \tag{5.99}$$

□

Now we are ready to prove Theorem 5.4. As usual, because of Corollary 4.16 we can ask

$$\mathbb{P}\left(\|F(0)-Y\|_2<R\right)\geq 1-\frac{\delta}{2}\quad\forall m\geq m_0 \tag{5.100}$$

for some $R=\sqrt{n\log(2n)}R_0$ with probability at least $1-\frac{\delta}{2}$.

At each temporal step, let us introduce two different evolutions of the same parameter vector Θ_t, one stochastic and one deterministic. More precisely, let us consider two stochastic processes Θ_t and $\tilde{\Theta}_t$ defined for $t\geq 0$ by

$$\begin{cases}\Theta_{t+1} &= \Theta_t-\eta g^{(t)}(\Theta_t)\\ \tilde{\Theta}_{t+1} &= \Theta_t-\eta\nabla_\Theta\mathcal{L}(\Theta_t)\end{cases}. \tag{5.101}$$

It is crucial to notice that the parameter vector appering in the RHS of the second equation is Θ_t, not $\tilde{\Theta}_t$. When we discussed the strategy to prove the theorem, we emphasized the necessity to bound the discrepancy between the loss function evalued at the new points Θ_{t+1} and $\tilde{\Theta}_{t+1}$. An upper bound can be computed starting from the Lipschitzness of the model. By Lemma 4.21,

$$|f(\Theta_1,x)-f(\Theta_2,x)|\leq 2|\Theta|\frac{|\mathcal{M}|}{N(m)}\|\Theta_1-\Theta_2\|_\infty. \tag{5.102}$$

This implies the following bound on the growth $\mathcal{L}(\Theta)$:

$$\begin{aligned}|\mathcal{L}(\Theta_1)-\mathcal{L}(\Theta_2)| &= \frac{1}{2n}\sum_{i=1}^{n}\left((f(\Theta_1,x^{(i)})-y^{(i)})^2-(f(\Theta_2,x^{(i)})-y^{(i)})^2\right)\\ &= \frac{1}{2n}\sum_{i=1}^{n}\left(f(\Theta_1,x^{(i)})-f(\Theta_2,x^{(i)})\right)\times\\ &\qquad\times\left(f(\Theta_1,x^{(i)})-y^{(i)}+f(\Theta_2,x^{(i)})-y^{(i)}\right)\\ &\leq \frac{1}{\sqrt{2n}}\|f(\Theta_1,X)-f(\Theta_2,X)\|_2\times\\ &\qquad\times\frac{1}{\sqrt{2n}}\left(\|f(\Theta_1,X)-Y\|_2+\|f(\Theta_2,X)-Y\|_2\right)\\ &\leq \sqrt{2}|\Theta|\frac{|\mathcal{M}|}{N(m)}\left(\sqrt{\mathcal{L}(\Theta_1)}+\sqrt{\mathcal{L}(\Theta_2)}\right)\|\Theta_1-\Theta_2\|_\infty.\end{aligned} \tag{5.103}$$

In particular,

$$\mathcal{L}(\Theta_1)\leq\mathcal{L}(\Theta_2)+\sqrt{2}|\Theta|\frac{|\mathcal{M}|}{N(m)}\sqrt{\mathcal{L}(\Theta_2)}\|\Theta_1-\Theta_2\|_\infty$$

$$+\sqrt{2}|\Theta|\frac{|\mathcal{M}|}{N(m)}\|\Theta_1-\Theta_2\|_\infty\sqrt{\mathcal{L}(\Theta_1)}. \tag{5.104}$$

So we can apply Lemma 5.13 with $x=\sqrt{\mathcal{L}(\Theta_1)}$ and

$$A=\mathcal{L}(\Theta_2),\qquad A'=\sqrt{2}|\Theta|\frac{|\mathcal{M}|}{N(m)}\sqrt{\mathcal{L}(\Theta_2)}\|\Theta_1-\Theta_2\|_\infty, \tag{5.105}$$

$$B=\sqrt{2}|\Theta|\frac{|\mathcal{M}|}{N(m)}\|\Theta_1-\Theta_2\|_\infty, \tag{5.106}$$

obtaining

$$\begin{aligned}\sqrt{\mathcal{L}(\Theta_1)}&\le\sqrt{\mathcal{L}(\Theta_2)}+\sqrt[4]{\mathcal{L}(\Theta_2)}\sqrt{\sqrt{2}|\Theta|\frac{|\mathcal{M}|}{N(m)}}\|\Theta_1-\Theta_2\|_\infty^{1/2}\\&\quad+\sqrt{2}|\Theta|\frac{|\mathcal{M}|}{N(m)}\|\Theta_1-\Theta_2\|_\infty\\&\le\left(\sqrt[4]{\mathcal{L}(\Theta_2)}+\sqrt{\sqrt{2}|\Theta|\frac{|\mathcal{M}|}{N(m)}}\|\Theta_1-\Theta_2\|_\infty^{1/2}\right)^2.\end{aligned} \tag{5.107}$$

The discrepancy between the loss function after a deterministic evolution of the parameters and the loss function after a stochastic evolution of the parameters is small when the outcome of the stochastic evolution is close to the result of the deterministic evolution. This can be done controlled by a concentration inequality. For any choice of $t\le T$ and $\xi_k\in\mathbb{R}$ and $\Theta_k\in\mathbb{R}^{|\Theta|}$,

$$\Theta_{t+1}-\tilde{\Theta}_{t+1}=\eta(\nabla_\Theta\mathcal{L}(\Theta_t)-g^{(t)}(\Theta_t)), \tag{5.108}$$

$$\mathbb{E}\left[\theta_i(t+1)-\tilde{\theta}_i(t+1)\;\middle|\;g^{(k)}(\Theta_k)=\xi_k\quad\forall\,k<t\right]=0, \tag{5.109}$$

$$\begin{aligned}&\mathrm{Var}\left[\theta_i(t+1)-\tilde{\theta}_i(t+1)\;\middle|\;g^{(k)}(\Theta_k)=\xi_k\quad\forall\,k<t\right]\\&\qquad=\eta^2\mathrm{Var}\left[g_i^{(t)}(\Theta_t)\;\middle|\;g^{(k)}(\Theta_k)=\xi_k\quad\forall\,k<t\right].\end{aligned} \tag{5.110}$$

In order to have a cleaner notation, from now on we will not explitictly write that the variance and the probability are contitioned on any possible outcome of the previous estimators, even though we will always assume this. By Chebyshev's inequality,

$$\mathbb{P}\left(|\theta_i(t+1)-\tilde{\theta}_i(t+1)|>k\eta\sqrt{\mathrm{Var}\left[g_i^{(t)}(\Theta_t)\right]}\right)\le\frac{1}{k^2}. \tag{5.111}$$

If we set

$$k=\pi\sqrt{\frac{|\Theta|}{3\delta}}(t+1), \tag{5.112}$$

then

$$\begin{aligned}&\mathbb{P}\left(|\theta_i(t+1)-\tilde{\theta}_i(t+1)|<\eta\pi\sqrt{\frac{|\Theta|}{3\delta}}(t+1)\sqrt{\mathrm{Var}\left[g_i^{(t)}(\Theta_t)\right]}\quad\forall\,t\le T,\forall i\right)\\&\qquad\ge1-\frac{\delta}{2}\frac{1}{|\Theta|}\sum_{i=1}^{Lm}\frac{6}{\pi^2}\sum_{t=0}^{T}\frac{1}{(t+1)^2}\ge1-\frac{\delta}{2}\frac{6}{\pi^2}\sum_{t=0}^{\infty}\frac{1}{(t+1)^2}\\&\qquad=1-\frac{\delta}{2}.\end{aligned} \tag{5.113}$$

So, if

$$\|\Theta_{t+1} - \tilde{\Theta}_{t+1}\|_\infty \leq \eta\pi\sqrt{\frac{|\Theta|}{3\delta}}(t+1)\sup_i \sqrt{\mathrm{Var}\left[g_i^{(t)}(\Theta_t)\right]}, \tag{5.114}$$

we can bound the loss function evalued at Θ_{t+1} by using (5.107)

$$\begin{aligned}\sqrt[4]{\mathcal{L}(\Theta_{t+1})} &\leq \sqrt[4]{\mathcal{L}(\tilde{\Theta}_{t+1})} + \sqrt{\sqrt{2}|\Theta|\frac{|\mathcal{M}|}{N(m)}}\|\Theta_{t+1} - \tilde{\Theta}_{t+1}\|_\infty^{1/2} \\ &\leq \sqrt[4]{\mathcal{L}(\tilde{\Theta}_{t+1})} + \sqrt[4]{\frac{2\eta^2\pi^2}{3\delta}|\Theta|^3\frac{|\mathcal{M}|^2}{N^2(m)}(t+1)^2\sup_i \mathrm{Var}\left[g_i^{(t)}(\Theta_t)\right]}.\end{aligned} \tag{5.115}$$

Now we prove the following bounds by induction on t:

$$\begin{cases}\|\Theta_t - \Theta_0\|_\infty \leq 2R_0\lambda_{\min}^K n\sqrt{\log(2n)}\dfrac{|\mathcal{M}|}{N_K(m)N(m)}\displaystyle\sum_{k=0}^{t-1}\left(1 - \frac{\eta_0\lambda_{\min}^K}{3}\right)^k \\ \mathcal{L}(\Theta_t) \leq \dfrac{1}{2}R_0^2\log(2n)\left(1 - \dfrac{\eta_0\lambda_{\min}^K}{3}\right)^{2t}\end{cases}. \tag{5.116}$$

If $t = 0$, it is clear that they hold. Now, let us assume that they hold for $t = t^*$.

Using the deterministic evolution

$$\tilde{\Theta}_{t^*+1} - \Theta_{t^*} = -\frac{\eta_0}{N_K(m)}\nabla_\Theta f(\Theta_{t^*}, X)\cdot(F(t) - Y) \tag{5.117}$$

and using Lemma 4.20

$$\begin{aligned}|\tilde{\theta}_i(t^*+1) - \theta_i(t^*)| &\leq \left|\frac{\eta_0}{N_K(m)}\partial_{\theta_i} f(\Theta_{t^*}, X)\cdot(F(t^*) - Y)\right| \\ &\leq \frac{\eta_0}{N_K(m)}\|\partial_{\theta_i} f(\Theta_{t^*}, X)\|_2\|F(t^*) - Y\|_2 \\ &\leq \frac{\eta_0}{N_K(m)}2\sqrt{n}\,\frac{|\mathcal{M}|}{N(m)}\sqrt{n\log(2n)}R_0\left(1 - \frac{1}{3}\eta_0\lambda_{\min}^K\right)^{t^*} \\ &= 2R_0\eta_0 n\sqrt{\log(2n)}\,\frac{|\mathcal{M}|}{N_K(m)N(m)}\left(1 - \frac{1}{3}\eta_0\lambda_{\min}^K\right)^{t^*}.\end{aligned} \tag{5.118}$$

By the triangle inequality, using the inductive hypothesis,

$$\begin{aligned}\|\tilde{\Theta}_{t^*+1} - \Theta_0\|_\infty &\leq \|\tilde{\Theta}_{t^*+1} - \Theta_{t^*}\|_\infty + \|\Theta_{t^*} - \Theta_0\|_\infty \\ &\leq 2R_0\eta_0 n\sqrt{\log(2n)}\,\frac{|\mathcal{M}|}{N_K(m)N(m)}\sum_{k=0}^{t^*}\left(1 - \frac{1}{3}\eta_0\lambda_{\min}^K\right)^k.\end{aligned} \tag{5.119}$$

In particular, both $\|\Theta_{t^*} - \Theta_0\|_\infty$ and $\|\tilde{\Theta}_{t^*+1} - \Theta_0\|_\infty$ are bounded by

$$\begin{aligned}&2R_0\eta_0 n\sqrt{\log(2n)}\,\frac{|\mathcal{M}|}{N_K(m)N(m)}\sum_{k=0}^{\infty}\left(1 - \frac{1}{3}\eta_0\lambda_{\min}^K\right)^k \\ &\qquad = \frac{6R_0}{N_K(m)\lambda_{\min}^K}n\sqrt{\log(2n)}\,\frac{|\mathcal{M}|}{N(m)} = \rho(m).\end{aligned} \tag{5.120}$$

Now, let us notice that

$$\lim_{m\to\infty} \frac{\Sigma_1|\mathcal{M}|^2|\mathcal{N}|}{N_K(m)N^2(m)}\rho(m) = \lim_{m\to\infty} \frac{6R_0}{\lambda^K_{\min}} n\sqrt{\log(2n)}\frac{\Sigma_1|\mathcal{M}|^3|\mathcal{N}|}{N_K^2(m)N^3(m)} = 0 \tag{5.121}$$

by hypothesis. We invoke[18] Lemma 5.11 with probablity to fail $\frac{\delta}{4}$ instead of δ and choosing the following parameters:

$$r(m) = \rho(m), \qquad \Theta' = \tilde{\Theta}_{t^*+1}, \qquad \Theta = \Theta_{t^*}, \qquad \zeta = \frac{2}{3}, \tag{5.122}$$

Therefore

$$\|f(\tilde{\Theta}_{t^*+1}, X) - Y\|_2 \leq \left(1 - \frac{2}{3}\eta_0\lambda^K_{\min}\right)\|f(\Theta_{t^*}, X) - Y\|_2, \tag{5.123}$$

which yields

$$\begin{aligned}\sqrt[4]{\mathcal{L}(\Theta_{t^*+1})} \leq& \sqrt{1 - \frac{2}{3}\eta_0\lambda^K_{\min}}\sqrt[4]{\mathcal{L}(\Theta_{t^*})} \\ &+ \sqrt[4]{\frac{2\eta^2\pi^2}{3\delta}|\Theta|^3\frac{|\mathcal{M}|^2}{N^2(m)}(t+1)^2\mathrm{Var}\left[g_i^{(t)}(\Theta_{t^*})\right]}.\end{aligned} \tag{5.124}$$

We recall that

$$\eta = \frac{n}{N_K(m)}\eta_0. \tag{5.125}$$

By hypothesis (5.19), the variance is sufficiently small:

$$\begin{aligned}\sqrt[4]{\mathcal{L}(\Theta_{t^*+1})} &\leq \sqrt{1 - \frac{2}{3}\eta_0\lambda^K_{\min}}\sqrt[4]{\mathcal{L}(\Theta_{t^*})} + \sqrt[4]{\frac{2\eta_0^2\pi^2}{3}c_0\eta_0^2\left(\lambda^K_{\min}\right)^4\mathcal{L}(\Theta_{t^*})} \\ &\leq \left(\sqrt{1 - \frac{2}{3}\eta_0\lambda^K_{\min}} + \eta_0\lambda^K_{\min}\sqrt[4]{\frac{2\pi^2}{3}\frac{1}{864\pi^2}}\right)\sqrt[4]{\mathcal{L}(\Theta_{t^*})} \\ &\leq \left(\sqrt{1 - \frac{2}{3}\eta_0\lambda^K_{\min}} + \frac{\eta_0\lambda^K_{\min}}{6}\right)\sqrt[4]{\mathcal{L}(\Theta_{t^*})}.\end{aligned} \tag{5.126}$$

Using Lemma 5.14 with $x/4 = \eta_0\lambda^K_{\min}/6$,

$$\begin{aligned}\sqrt[4]{\mathcal{L}(\Theta_{t^*+1})} &\leq \left(\sqrt{1 - \frac{2}{3}\eta_0\lambda^K_{\min}} + \sqrt{1 - \frac{\eta_0\lambda^K_{\min}}{3}} - \sqrt{1 - \frac{2}{3}\eta_0\lambda^K_{\min}}\right)\sqrt[4]{\mathcal{L}(\Theta_{t^*})} \\ &= \sqrt{1 - \frac{\eta_0\lambda^K_{\min}}{3}}\sqrt[4]{\mathcal{L}(\Theta_{t^*})},\end{aligned} \tag{5.127}$$

whence

$$\mathcal{L}(\Theta_{t^*+1}) \leq \left(1 - \frac{\eta_0\lambda^K_{\min}}{3}\right)^2\mathcal{L}(\Theta_{t^*}). \tag{5.128}$$

Then, using the inductive hypothesis,

$$\mathcal{L}(\Theta_{t^*+1}) \leq \frac{R_0^2}{2}\log(2n)\left(1 - \frac{\eta_0\lambda^K_{\min}}{3}\right)^{2(t^*+1)}. \tag{5.129}$$

[18] Since $r(m)$ and ζ are always the same in the proof by induction, the lemma has to be claimed once.

So, both the bounds (5.116) hold for $t = t^* + 1$. By induction, they are valid for any $t \in \mathbb{N}$. Using Lemma 4.20,

$$
\begin{aligned}
|\tilde{\theta}_i(t+1) - \theta_i(t)| = \eta |\partial_{\theta_i} \mathcal{L}(\Theta_t)| &= \frac{\eta}{n} \left| \sum_{j=1}^{n} \partial_{\theta_i} f(\Theta_t, x^{(j)}) \big(f(\Theta_t, x^{(j)}) - y^{(j)} \big) \right| \\
&\leq \frac{\sqrt{2}\eta}{\sqrt{n}} \|\partial_{\theta_i} f(\Theta, X)\|_2 \frac{1}{\sqrt{2n}} \|F(t) - Y\|_2 \\
&= 2\sqrt{2}\eta \frac{|\mathcal{M}|}{N(m)} \sqrt{\mathcal{L}(\Theta_t)}. \qquad (5.130)
\end{aligned}
$$

By the triangle inequality,

$$
\begin{aligned}
\|\Theta_{t+1} - \Theta_t\|_\infty &\leq \|\Theta_{t+1} - \tilde{\Theta}_{t+1}\|_\infty + \|\tilde{\Theta}_{t+1} - \Theta_t\|_\infty \\
&\leq \eta\pi \sqrt{\frac{|\Theta|}{3\delta}} (t+1) \sqrt{\mathrm{Var}\left[g_i^{(t)}(\Theta_t) \right]} + 2\sqrt{2}\eta \frac{|\mathcal{M}|}{N(m)} \sqrt{\mathcal{L}(\Theta_t)} \\
&\leq \eta\pi \sqrt{\frac{|\Theta|}{3\delta}} (t+1) \sqrt{c_0}\eta_0 \frac{\left(\lambda^K_{\min}\right)^2}{n} \frac{N_K(m) N(m)}{|\mathcal{M}||\Theta|^{3/2}} \frac{\sqrt{\delta}}{(t+1)} \sqrt{\mathcal{L}(\Theta_t)} \\
&\quad + 2\sqrt{2}\eta \frac{|\mathcal{M}|}{N(m)} \sqrt{\mathcal{L}(\Theta_t)} \\
&\leq \eta_0 \frac{R_0}{\sqrt{2}} \sqrt{\log(2n)} \times \\
&\quad \times \left(\eta_0 \frac{\pi\sqrt{c_0}}{\sqrt{3}} \left(\lambda^K_{\min}\right)^2 \frac{N(m)}{|\mathcal{M}||\Theta|} + 2\sqrt{2}\, n \frac{|\mathcal{M}|}{N_K(m) N(m)} \right) \left(1 - \frac{\eta_0 \lambda^K_{\min}}{3} \right)^t . \qquad (5.131)
\end{aligned}
$$

Hence, using again the triangle inequality,

$$
\begin{aligned}
\|\Theta_T - \Theta_0\|_\infty &\leq \sum_{t=0}^{T-1} \|\Theta_{t+1} - \Theta_t\|_\infty \\
&\leq \eta_0 R_0 \sqrt{\log(2n)} \left(\eta_0 \frac{\pi\sqrt{c_0}}{\sqrt{6}} \left(\lambda^K_{\min}\right)^2 \frac{N(m)}{|\mathcal{M}||\Theta|} + 2n \frac{|\mathcal{M}|}{N_K(m) N(m)} \right) \times \\
&\quad \times \sum_{t=0}^{\infty} \left(1 - \frac{\eta_0 \lambda^K_{\min}}{3} \right)^t \\
&= \frac{3 R_0 \sqrt{\log(2n)}}{\lambda^K_{\min}} \left(\eta_0 \frac{\pi\sqrt{c_0}}{\sqrt{6}} \left(\lambda^K_{\min}\right)^2 \frac{N(m)}{|\mathcal{M}||\Theta|} + 2n \frac{|\mathcal{M}|}{N_K(m) N(m)} \right) \\
&= \sqrt{\log(2n)} \left(R_1(\delta) \eta_0 \lambda^K_{\min} \frac{N(m)}{|\mathcal{M}||\Theta|} + R_2(\delta) \frac{n}{\lambda^K_{\min}} \frac{|\mathcal{M}|}{N_K(m) N(m)} \right), \qquad (5.132)
\end{aligned}
$$

where we recalled the dependence $R_0 = R_0(\delta)$. This proves (5.21). Finally, we use Theorem 4.23 to give the upper bound

$$
\begin{aligned}
&|f(\Theta, x) - f^{\mathrm{lin}}(\Theta, x)| \\
&\quad \leq \frac{Lm|\mathcal{M}|^2|\mathcal{N}|}{N(m)} \log(2n) \left(R_1(\delta) \eta_0 \lambda^K_{\min} \frac{N(m)}{|\mathcal{M}||\Theta|} + R_2(\delta) \frac{n}{\lambda^K_{\min}} \frac{|\mathcal{M}|}{N_K(m) N(m)} \right)^2 \\
&\quad \leq 2 R_1^2(\delta) \log(2n) \eta_0^2 \left(\lambda^K_{\min}\right)^2 \frac{|\mathcal{N}| N(m)}{Lm} + 2 \frac{R_2^2(\delta)}{\left(\lambda^K_{\min}\right)^2} n^2 \log(2n) \frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m) N^3(m)}, \qquad (5.133)
\end{aligned}
$$

where in the last line we have used that $(x+y)^2 \leq 2x^2 + 2y^2$. And this concludes the proof.

5.6 Proof of Lemma 5.6

We need to introduce some definitions in order to deal with the discretization of time and with the noise. Let us consider the case in which the gradient of the cost is exactly known and let us define $r(t,x)$ as

$$f(\Theta_{t+1},x) =: f(\Theta_t,x) + \nabla_\Theta f(\Theta_t,x)^T(\Theta_{t+1}-\Theta_t) + r(t,x), \tag{5.134}$$

so that the discrete evolution reads

$$\begin{aligned}\Delta f(\Theta_t,x) := f(\Theta_{t+1},x) - f(\Theta_t,x) &= \nabla_\Theta f(\Theta_t,x)^T(\Theta_{t+1}-\Theta_t) + r(t,x)\\ &= -\eta\nabla_\Theta f(\Theta_t,x)^T\nabla_\Theta\mathcal{L}(\Theta_t) + r(t,x)\\ &= -\eta_0\hat{K}_{\Theta_t}(x,X^T)(F(t)-Y) + r(t,x).\end{aligned} \tag{5.135}$$

Let us now generalize the previous equation in the presence of the noise due the finite number of measurements; as before, we introduce $r(t,x)$ as

$$f(\Theta_{t+1},x) =: f(\Theta_t,x) + \nabla_\Theta f(\Theta_t,x)^T(\Theta_{t+1}-\Theta_t) + r(t,x), \tag{5.136}$$

but we also define

$$g^{(t)}(\Theta_t) =: \nabla\mathcal{L}(\Theta_t) + \epsilon(t) \qquad \text{and} \qquad \delta(t,x) =: -\eta\nabla_\Theta f(\Theta_t,x)^T\epsilon(t) \tag{5.137}$$

so that

$$\Theta_{t+1}-\Theta_t = -\eta g^{(t)}(\Theta_t) = -\eta\big(\nabla_\Theta\mathcal{L}(\Theta_t)+\epsilon(t)\big) \tag{5.138}$$

whence

$$\begin{aligned}\Delta f(\Theta_t,x) := f(\Theta_{t+1},x) - f(\Theta_t,x) &= \nabla_\Theta f(\Theta_t,x)^T(\Theta_{t+1}-\Theta_t) + r(t,x)\\ &= -\eta\nabla_\Theta f(\Theta_t,x)^T\big(\nabla_\Theta\mathcal{L}(\Theta_t)+\epsilon(t)\big) + r(t,x)\\ &= -\eta_0\hat{K}_{\Theta_t}(x,X^T)(F(t)-Y) + r(t,x) + \delta(t,x),\end{aligned} \tag{5.139}$$

therefore the noiseless case can be recovered from (5.139) by setting $\delta(t) = 0$. We extend the time parameter t to semi-integers as follows

$$\begin{aligned}f(\Theta_{t+1/2},x) :&= \frac{f(\Theta_{t+1},x)+f(\Theta_t,x)}{2}\\ &= \frac{1}{2}\big(2f(\Theta_t,x)+\Delta f(\Theta_t,x)\big)\\ &= f(\Theta_t,x) - \frac{\eta_0}{2}\hat{K}_{\Theta_t}(x,X^T)(F(t)-Y) + \frac{1}{2}(r(t,x)+\delta(t,x)).\end{aligned} \tag{5.140}$$

This compact notation will be useful in the upcoming computations. It is simple to verify that

$$\begin{aligned}f^{\text{lin}}(\Theta_{t+1/2},x) :&= \frac{f^{\text{lin}}(\Theta_{t+1},x)+f^{\text{lin}}(\Theta_t,x)}{2}\\ &= f^{\text{lin}}(\Theta_t,x) - \frac{\eta_0}{2}\hat{K}_{\Theta_0}(x,X^T)(F^{\text{lin}}(t)-Y).\end{aligned} \tag{5.141}$$

Later we will also use that, for any function $g(t)$, the following identity holds

$$\begin{aligned}\frac{1}{2}\Delta g^2(t) &= \frac{1}{2}\big(g^2(t+1)-g^2(t)\big)\\ &= \frac{1}{2}\big(g(t+1)+g(t)\big)\big(g(t+1)-g(t)\big) = g(t+1/2)\Delta g(t),\end{aligned} \tag{5.142}$$

where we used the previous convention on semi-integers time values.

The following lemma is a crucial ingredient to prove Lemma 5.6.

Lemma 5.15 (Bounding the discrepancy on the examples, discrete time case) *Let us assume a condition on the variance of the estimator of the gradient stronger than (5.19) by a factor $\xi(m)$: for any choice of $t \leq T$ and $\xi_k \in \mathbb{R}$ and $\Theta_k \in \mathbb{R}^{|\Theta|}$,*

$$\begin{aligned}
&\mathrm{Var}\Big[g_i^{(t)}(\Theta) \;\Big|\; g^{(k)}(\Theta_k) = \xi_k \quad \forall\, k < t\Big] \\
&\qquad \leq c_0 \eta_0^2 \frac{\left(\lambda_{\min}^K\right)^4}{n^2} \frac{N_K^2(m) N^2(m)}{|\mathcal{M}|^2 |\Theta|^3} \xi(m) \frac{\delta/4}{(t+1)^2} \mathcal{L}(\Theta),
\end{aligned} \tag{5.143}$$

where $c_0 = \frac{1}{8 6 4 \pi^2}$. Then, for any $\delta > 0$, there exists C, C' and $m_0 \in \mathbb{N}$ such that, for each $m \geq m_0$

$$\|F(t) - F^{\mathrm{lin}}(t)\|_2 \leq \frac{C}{(\lambda_{\min}^K)^2} n^2 \sqrt{n} \log(2n) \frac{Lm|\mathcal{M}|^4 |\mathcal{N}|^3}{N^3(m)} + \Delta(m) \tag{5.144}$$

with probability at least $1 - \delta$, where

$$\Delta(m) = C' \eta_0 \lambda_{\min}^K \sqrt{n \log(2n)} \sqrt{\xi(m)} \tag{5.145}$$

Proof We will use the compact notation mentioned above:

$$F(t) = f(\Theta_t, X), \quad F^{\mathrm{lin}}(t) = f^{\mathrm{lin}}(\Theta_t^{\mathrm{lin}}, X) \tag{5.146}$$

$$F(t+1/2) = \frac{f(\Theta_{t+1}, X) + f(\Theta_t, X)}{2}, \quad F^{\mathrm{lin}}(t+1/2) = \frac{f^{\mathrm{lin}}(\Theta_{t+1}^{\mathrm{lin}}, X) + f(\Theta_t^{\mathrm{lin}}, X)}{2} \tag{5.147}$$

We recall that $\hat{K}_{\Theta_t} := \hat{K}_{\Theta_t}(X, X^T)$.

$$\begin{aligned}
&\frac{1}{2}\Delta ||F(t) - F^{\mathrm{lin}}(t)||_2^2 \\
&\quad = \frac{1}{2}\left(||F(t+1) - F^{\mathrm{lin}}(t+1)||_2^2 - ||F(t) - F^{\mathrm{lin}}(t)||_2^2\right) \\
&\quad = \Big(F(t+1) - F(t) - F^{\mathrm{lin}}(t+1) + F^{\mathrm{lin}}(t)\Big)^T \Big(F(t+1/2) - F^{\mathrm{lin}}(t+1/2)\Big) \\
&\quad = \Big(\Delta F(t) - \Delta F^{\mathrm{lin}}(t)\Big)^T \Big(F(t+1/2) - F^{\mathrm{lin}}(t+1/2)\Big) \\
&\quad = -\eta_0 \Big(\hat{K}_{\Theta_t}(F(t) - Y) - \hat{K}_{\Theta_0}(F^{\mathrm{lin}}(t) - Y)\Big)^T \Big(F(t+1/2) - F^{\mathrm{lin}}(t+1/2)\Big) \\
&\qquad + \big(r(t,X) + \delta(t,X)\big)^T \Big(F(t+1/2) - F^{\mathrm{lin}}(t+1/2)\Big) \\
&\quad = -\eta_0 \Big((\hat{K}_{\Theta_t} - \hat{K}_{\Theta_0})(F(t) - Y) - \hat{K}_{\Theta_0}(F^{\mathrm{lin}}(t) - F(t))\Big)^T \times \\
&\qquad \times \Big(F(t+1/2) - F^{\mathrm{lin}}(t+1/2)\Big) \\
&\qquad + \big(r(t,X) + \delta(t,X)\big)^T \Big(F(t+1/2) - F^{\mathrm{lin}}(t+1/2)\Big) \\
&\quad = -\eta_0 (F(t) - Y)^T (\hat{K}_{\Theta_t} - \hat{K}_{\Theta_0}) \Big(F(t+1/2) - F^{\mathrm{lin}}(t+1/2)\Big) \\
&\qquad + \eta_0 (F^{\mathrm{lin}}(t) - F(t))^T \hat{K}_{\Theta_0} \Big(F(t+1/2) - F^{\mathrm{lin}}(t+1/2)\Big) \\
&\qquad + \big(r(t,X) + \delta(t,X)\big)^T \Big(F(t+1/2) - F^{\mathrm{lin}}(t+1/2)\Big) \\
&\quad = A(t) + B(t) + C(t)
\end{aligned} \tag{5.148}$$

where

$$A(t) := -\eta_0 (F(t) - Y)^T (\hat{K}_{\Theta_t} - \hat{K}_{\Theta_0}) \Big(F(t+1/2) - F^{\mathrm{lin}}(t+1/2)\Big) \tag{5.149}$$

$$B(t) := \eta_0 (F^{\mathrm{lin}}(t) - F(t))^T \hat{K}_{\Theta_0} \Big(F(t+1/2) - F^{\mathrm{lin}}(t+1/2)\Big) \tag{5.150}$$

$$C(t) := \big(r(t,X) + \delta(t,X)\big)^T \Big(F(t+1/2) - F^{\mathrm{lin}}(t+1/2)\Big) \tag{5.151}$$

Now, in the absence of statistical noise thanks to Theorem 5.1 – which we invoke with probability at least $1 - \frac{\delta}{4}$ – we have (5.8). In the presence of statistical noise, we invoke Theorem 5.4 with probability $1 - \frac{\delta}{4}$, so that (5.41) ensures that (5.19) is verified. Using the analogue of Corollary 4.11 in the discrete time setting (the proof would

be identical) with probability at least $1-\frac{\delta}{4}$, we have the following bound, which holds with probability at least $1-\frac{\delta}{2}$:

$$
\begin{aligned}
|A(t)| &\leq \eta_0 \|F(t)-Y\|_2 \|\hat{K}_{\Theta_t}-\hat{K}_{\Theta_0}\|_{\mathcal{L}} \left\|F(t+1/2)-F^{\text{lin}}(t+1/2)\right\|_2 \\
&\leq \eta_0 R_0 \sqrt{n\log(2n)} \left(1-\frac{\eta_0\lambda_{\min}^K}{3}\right)^t \|\hat{K}_{\Theta_t}-\hat{K}_{\Theta_0}\|_F \times \\
&\quad \times \frac{1}{2}\left\|F(t+1)-F^{\text{lin}}(t+1)+F(t)-F^{\text{lin}}(t)\right\|_2 \\
&\leq \eta_0 R_0 \sqrt{n\log(2n)} \left(1-\frac{\eta_0\lambda_{\min}^K}{3}\right)^t \frac{R_2}{\lambda_{\min}^K} n^2\sqrt{\log(2n)} \frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m)N^3(m)} \times \\
&\quad \times \frac{1}{2}\left(\left\|F(t+1)-F^{\text{lin}}(t+1)\right\|+\left\|F(t)-F^{\text{lin}}(t)\right\|_2\right) \\
&= \frac{R_0R_2\eta_0}{\lambda_{\min}^K} n^2\sqrt{n}\log(2n)\frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m)N^3(m)}\left(1-\frac{\eta_0\lambda_{\min}^K}{3}\right)^t \times \\
&\quad \times \frac{1}{2}\left(\left\|F(t+1)-F^{\text{lin}}(t+1)\right\|+\left\|F(t)-F^{\text{lin}}(t)\right\|_2\right).
\end{aligned} \tag{5.152}
$$

Using (5.140) and (5.141),

$$
\begin{aligned}
B(t) &= \eta_0(F^{\text{lin}}(t)-F(t))^T\hat{K}_{\Theta_0}\left(F(t+1/2)-F^{\text{lin}}(t+1/2)\right) \\
&= \eta_0(F^{\text{lin}}(t)-F(t))^T\hat{K}_{\Theta_0}\Big(F(t)-\frac{\eta_0}{2}\hat{K}_{\Theta_t}(F(t)-Y) \\
&\qquad\qquad -F^{\text{lin}}(t)+\frac{\eta_0}{2}\hat{K}_{\Theta_0}(F^{\text{lin}}(t)-Y)\Big) \\
&\quad +\frac{\eta_0}{2}(F^{\text{lin}}(t)-F(t))^T\hat{K}_{\Theta_0}(r(t,X)+\delta(t,X)) \\
&= -\eta_0(F^{\text{lin}}(t)-F(t))^T\hat{K}_{\Theta_0}(F^{\text{lin}}(t)-F(t)) \\
&\quad -\frac{\eta_0^2}{2}(F^{\text{lin}}(t)-F(t))^T\hat{K}_{\Theta_0}\left(\hat{K}_{\Theta_t}(F(t)-Y)-\hat{K}_{\Theta_0}(F^{\text{lin}}(t)-Y)\right) \\
&\quad +\frac{\eta_0}{2}(F^{\text{lin}}(t)-F(t))^T\hat{K}_{\Theta_0}(r(t,X)+\delta(t,X)) \\
&= -\eta_0(F^{\text{lin}}(t)-F(t))^T\hat{K}_{\Theta_0}(F^{\text{lin}}(t)-F(t)) \\
&\quad -\frac{\eta_0^2}{2}(F^{\text{lin}}(t)-F(t))^T\hat{K}_{\Theta_0}\left(\hat{K}_{\Theta_t}(F(t)-Y)-\hat{K}_{\Theta_0}(F^{\text{lin}}(t)-Y)\right) \\
&\quad +\frac{\eta_0}{2}(F^{\text{lin}}(t)-F(t))^T\hat{K}_{\Theta_0}(r(t,X)+\delta(t,X)) \\
&= -\eta_0(F^{\text{lin}}(t)-F(t))^T\hat{K}_{\Theta_0}(F^{\text{lin}}(t)-F(t)) \\
&\quad -\frac{\eta_0^2}{2}(F^{\text{lin}}(t)-F(t))^T\hat{K}_{\Theta_0}\left((\hat{K}_{\Theta_t}-\hat{K}_{\Theta_0})(F(t)-Y)-\hat{K}_{\Theta_0}(F^{\text{lin}}(t)-F(t))\right) \\
&\quad +\frac{\eta_0}{2}(F^{\text{lin}}(t)-F(t))^T\hat{K}_{\Theta_0}(r(t,X)+\delta(t,X)) \\
&= (F^{\text{lin}}(t)-F(t))^T\left(-\eta_0\hat{K}_{\Theta_0}+\frac{\eta_0^2}{2}\hat{K}_{\Theta_0}^2\right)(F^{\text{lin}}(t)-F(t)) \\
&\quad -\frac{\eta_0^2}{2}(F^{\text{lin}}(t)-F(t))^T\hat{K}_{\Theta_0}(\hat{K}_{\Theta_t}-\hat{K}_{\Theta_0})(F(t)-Y) \\
&\quad +\frac{\eta_0}{2}(F^{\text{lin}}(t)-F(t))^T\hat{K}_{\Theta_0}(r(t,X)+\delta(t,X)).
\end{aligned} \tag{5.153}
$$

We recall that, in the discrete time setting, we always require that

$$
\eta_0 < \frac{2}{\lambda_{\min}^K+\lambda_{\max}^K}, \tag{5.154}
$$

in particular

$$
\eta_0 < \frac{2}{\lambda_{\max}^K}, \tag{5.155}
$$

therefore it exists $\rho>1$ such that

$$
\eta_0 = \frac{1}{\rho}\frac{2}{\lambda_{\max}^K}. \tag{5.156}
$$

The spectrum of $\hat{K}_{\Theta_0}$ converges to the spectrum of $\bar{K}$, so, if m is large enough,

$$\mathrm{Spec}(\hat{K}_{\Theta_0}) \subseteq [0, \rho\lambda_{\max}] = [0, 2/\eta_0]. \tag{5.157}$$

Now, we notice that

$$\phi(\xi) = \frac{1}{2}\xi^2 - \xi \leq 0 \quad \iff \quad \xi \in [0,2] \tag{5.158}$$

hence, for any ξ of the form

$$\xi = \eta_0\lambda \qquad \lambda \in \mathrm{Spec}(\hat{K}_{\Theta_0}), \tag{5.159}$$

we have $\phi(\xi) \leq 0$, so the matrix

$$M = -\eta_0\hat{K}_{\Theta_0} + \frac{\eta_0^2}{2}\hat{K}_{\Theta_0}^2 \tag{5.160}$$

is negative semidefinite. Whence,

$$\begin{aligned} B(t) \leq & -\frac{\eta_0^2}{2}(F^{\mathrm{lin}}(t) - F(t))^T\hat{K}_{\Theta_0}(\hat{K}_{\Theta_t} - \hat{K}_{\Theta_0})(F(t) - Y) \\ & + \frac{\eta_0}{2}(F^{\mathrm{lin}}(t) - F(t))^T\hat{K}_{\Theta_0}(r(t,X) + \delta(t,X)). \end{aligned} \tag{5.161}$$

We recall that we just required the maximal eigenvalue of $\hat{K}_{\Theta_0}$ to be smaller than $2/\eta_0$, so

$$\|\hat{K}_{\Theta_0}\|_{\mathcal{L}} \leq \frac{2}{\eta_0}, \tag{5.162}$$

which will be useful soon. Indeed,

$$\begin{aligned} |B(t)| \leq & \frac{\eta_0^2}{2}\|F^{\mathrm{lin}}(t) - F(t)\|_2\|\hat{K}_{\Theta_0}\|_{\mathcal{L}}\|\hat{K}_{\Theta_t} - \hat{K}_{\Theta_0}\|_{\mathcal{L}}\|F(t) - Y\|_2 \\ & + \frac{\eta_0}{2}\|F^{\mathrm{lin}}(t) - F(t)\|_2\|\hat{K}_{\Theta_0}\|_{\mathcal{L}}\|r(t,X) + \delta(t,X)\|_2 \\ \leq & \|F^{\mathrm{lin}}(t) - F(t)\|_2\left(\frac{\eta_0^2}{2}\frac{2}{\eta_0}\|\hat{K}_{\Theta_t} - \hat{K}_{\Theta_0}\|_F R_0\sqrt{n\log(2n)}\left(1 - \frac{\eta_0\lambda_{\min}^K}{3}\right)^t\right. \\ & \left. + \frac{\eta_0}{2}\frac{2}{\eta_0}\|r(t,X) + \delta(t,X)\|_2\right) \\ \leq & \|F^{\mathrm{lin}}(t) - F(t)\|_2\left(\frac{\eta_0^2}{2}\frac{2}{\eta_0}\frac{R_2}{\lambda_{\min}^K}n^2\sqrt{\log(2n)}\frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m)N^3(m)}R_0\sqrt{n\log(2n)}\times\right. \\ & \left.\times\left(1 - \frac{\eta_0\lambda_{\min}^K}{3}\right)^t + \frac{\eta_0}{2}\frac{2}{\eta_0}\|r(t,X) + \delta(t,X)\|_2\right) \\ \leq & \|F^{\mathrm{lin}}(t) - F(t)\|_2\left(\frac{R_0R_2\eta_0}{\lambda_{\min}^K}n^2\sqrt{n}\log(2n)\frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m)N^3(m)}\left(1 - \frac{\eta_0\lambda_{\min}^K}{3}\right)^t\right. \\ & \left. + \|r(t,X)\|_2 + \|\delta(t,X)\|_2\right) \end{aligned} \tag{5.163}$$

Finally,

$$\begin{aligned} |C(t)| \leq & \|r(t,X) + \delta(t,X)\|_2\left\|F(t+1/2) - F^{\mathrm{lin}}(t+1/2)\right\|_2 \\ \leq & (\|r(t,X)\|_2 + \|\delta(t,X)\|_2) \times \frac{1}{2}\left(\left\|F(t+1) - F^{\mathrm{lin}}(t+1)\right\| + \left\|F(t) - F^{\mathrm{lin}}(t)\right\|_2\right) \end{aligned} \tag{5.164}$$

Now, as we discussed above, we use that

$$\frac{1}{2}\Delta g^2(t) = g(t+1/2)\Delta g(t) \tag{5.165}$$

with

$$g(t) = \|F(t) - F^{\mathrm{lin}}(t)\|_2 \tag{5.166}$$

so that

$$\begin{aligned} & \frac{1}{2}\big(\|F(t+1) - F^{\mathrm{lin}}(t+1)\|_2 + \|F(t) - F^{\mathrm{lin}}(t)\|_2\big)\Delta\|F(t) - F^{\mathrm{lin}}(t)\|_2 \\ & \qquad = \frac{1}{2}\Delta\|F(t) - F^{\mathrm{lin}}(t)\|_2^2 = A(t) + B(t) + C(t). \end{aligned} \tag{5.167}$$

Hence,

$$\left|\Delta\|F(t)-F^{\text{lin}}(t)\|_2\right| \leq \frac{|A(t)|+|B(t)|+|C(t)|}{\frac{1}{2}\left(\|F(t+1)-F^{\text{lin}}(t+1)\|_2+\|F(t)-F^{\text{lin}}(t)\|_2\right)}. \tag{5.168}$$

Using our bounds and recalling that, for any $a, b > 0$,

$$\frac{a}{a+b} \leq 1, \tag{5.169}$$

we find that

$$\begin{aligned}\left|\Delta\|F(t)-F^{\text{lin}}(t)\|_2\right| \leq & \left(\frac{R_0 R_2 \eta_0}{\lambda_{\min}^K} n^2 \sqrt{n}\log(2n)\frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m)N^3(m)}\left(1-\frac{\eta_0\lambda_{\min}^K}{3}\right)^t + \right. \\ & \left. + \|r(t,X)\|_2 + \|\delta(t,X)\|_2\right)\times \\ & \times\left(1+\frac{\|F(t)-F^{\text{lin}}(t)\|_2}{\frac{1}{2}\left(\|F(t+1)-F^{\text{lin}}(t+1)\|_2+\|F(t)-F^{\text{lin}}(t)\|_2\right)}\right) \\ \leq & \, 3\left(\frac{R_0 R_2 \eta_0}{\lambda_{\min}^K} n^2 \sqrt{n}\log(2n)\frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m)N^3(m)}\left(1-\frac{\eta_0\lambda_{\min}^K}{3}\right)^t + \right. \\ & \left. + \|r(t,X)\|_2 + \|\delta(t,X)\|_2\right)\end{aligned} \tag{5.170}$$

Due to Theorem 4.23,

$$|r(t,x)| \leq \frac{Lm|\mathcal{M}|^2|\mathcal{N}|}{N(m)}\|\Theta_{t+1}-\Theta_t\|_\infty^2 \tag{5.171}$$

so, recalling (5.131)

$$\begin{aligned}\|\Theta_{t+1}-\Theta_t\|_\infty \leq & \, \eta_0\frac{R_0}{\sqrt{2}}\sqrt{\log(2n)}\times \\ & \times\left(\eta_0\frac{\pi\sqrt{c_0}}{\sqrt{3}}\left(\lambda_{\min}^K\right)^2\frac{N(m)}{|\mathcal{M}||\Theta|}+2\sqrt{2}\,n\frac{|\mathcal{M}|}{N_K(m)N(m)}\right)\left(1-\frac{\eta_0\lambda_{\min}^K}{3}\right)^t,\end{aligned} \tag{5.172}$$

we can bound

$$\begin{aligned}\sum_{t=0}^{\infty}|r(t,x)| \leq & \, 2\eta_0^2\frac{R_0^2}{2}\log(2n)\times \\ & \times\left(\eta_0^2\frac{\pi^2 c_0}{3}\left(\lambda_{\min}^K\right)^2\frac{N(m)|\mathcal{N}|}{Lm}+8\,n^2\frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m)N^3(m)}\right)\frac{3}{2\eta_0\lambda_{\min}^K}, \\ \leq & \, \frac{3R_0^2}{\lambda_{\min}^K}\eta_0\log(2n)\left(\eta_0^2\frac{\pi^2 c_0}{3}\left(2\lambda_{\min}^K\right)^2\frac{N(m)|\mathcal{N}|}{Lm}+8\,n^2\frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m)N^3(m)}\right),\end{aligned} \tag{5.173}$$

where we used that

$$\sum_{t=0}^{\infty}\left(1-\frac{\eta_0\lambda_{\min}^K}{3}\right)^{2t} = \frac{1}{\frac{2\eta_0\lambda_{\min}^K}{3}-\left(\frac{\eta_0\lambda_{\min}^K}{3}\right)^2} \leq \frac{3}{2\eta_0\lambda_{\min}^K}. \tag{5.174}$$

Let us claim that the variance of the statistical noise is small enough

$$\sum_{t=0}^{\infty}3\|\delta(t,X)\|_2 \leq \Delta(m) \tag{5.175}$$

with high probability, where $\Delta(m) \to 0$ fast enough when $m \to \infty$. More precisely, we have previously invoked Theorem 5.4, so we can assume that (5.114) holds, i.e.

$$\begin{aligned}\|\eta(\nabla_\Theta\mathcal{L}(\Theta_t)-g^{(t)}(\Theta_t))\|_\infty =: & \, \|\Theta_{t+1}-\tilde{\Theta}_{t+1}\|_\infty \\ & \leq \eta\pi\sqrt{\frac{|\Theta|}{3\delta}}(t+1)\sup_i\sqrt{\text{Var}\left[g_i^{(t)}(\Theta_t)\right]}.\end{aligned} \tag{5.176}$$

Recalling that

$$g^{(t)}(\Theta_t) = \nabla\mathcal{L}(\Theta_t)+\epsilon(t) \qquad \text{and} \qquad \delta(t,x) = -\eta\nabla_\Theta f(\Theta_t,x)^T\epsilon(t) \tag{5.177}$$

and using (5.20) in (5.41) to write explicitly

$$\begin{aligned}\|\epsilon(t)\|_\infty &= \|\nabla_\Theta \mathcal{L}(\Theta_t) - g^{(t)}(\Theta_t)\|_\infty \\ &\leq \pi\sqrt{\frac{|\Theta|}{3\delta}}(t+1)\sup_i \sqrt{\mathrm{Var}\left[g_i^{(t)}(\Theta_t)\right]} \\ &\leq \pi\sqrt{\frac{|\Theta|}{3\delta}}(t+1)\Bigg(c_0\eta_0^2\frac{\left(\lambda_{\min}^K\right)^4}{n^2}\frac{N_K^2(m)N^2(m)}{|\mathcal{M}|^2|\Theta|^3}\times \\ &\qquad \times \frac{\delta/4}{(t+1)^2}\frac{R_0^2}{2}\log(2n)\left(1-\frac{1}{3}\eta_0\lambda_{\min}^K\right)^{2t}\Bigg)^{1/2}\sqrt{\xi(m)}, \end{aligned} \tag{5.178}$$

we finally know that, with the estimate (4.117),

$$\begin{aligned}\|\delta(t,X)\|_2 &\leq \eta\sqrt{n}\sup_x \|\nabla_\Theta f(\Theta_t,x)\|_1\|\epsilon(t)\|_\infty \\ &\leq \eta\sqrt{n}\left(2L\frac{m}{N(m)}|\mathcal{M}|\right)\pi\sqrt{\frac{|\Theta|}{3\delta}}(t+1)\sqrt{\xi(m)}\times \\ &\quad\times\left(c_0\eta_0^2\frac{\left(\lambda_{\min}^K\right)^4}{n^2}\frac{N_K^2(m)N^2(m)}{|\mathcal{M}|^2|\Theta|^3}\frac{\delta/4}{(t+1)^2}\frac{R_0^2}{2}\log(2n)\left(1-\frac{1}{3}\lambda_{\min}^K\right)^{2t}\right)^{1/2} \\ &\leq R_0\sqrt{c_0}\sqrt{\log(2n)}\frac{\left(\lambda_{\min}^K\right)^2}{\sqrt{n}}\left(N_K(m)\eta\eta_0\right)\pi\sqrt{\frac{1}{6}}\left(1-\frac{1}{3}\eta_0\lambda_{\min}^K\right)^t\sqrt{\xi(m)} \\ &\leq R_0\pi\sqrt{\frac{c_0}{6}}\sqrt{n\log(2n)}\left(\eta_0\lambda_{\min}^K\right)^2\left(1-\frac{1}{3}\eta_0\lambda_{\min}^K\right)^t\sqrt{\xi(m)} \end{aligned} \tag{5.179}$$

where we recalled that $N_K(m)\eta = n\eta_0$. Then

$$\begin{aligned}\sum_{t=0}^\infty 3\|\delta(t,X)\|_2 &\leq \frac{3\pi R_0}{\frac{1}{3}\eta_0\lambda_{\min}^K}\sqrt{\frac{c_0}{6}}\sqrt{n\log(2n)}\left(\eta_0\lambda_{\min}^K\right)^2\sqrt{\xi(m)} \\ &\leq 9\pi R_0\sqrt{\frac{c_0}{6}}\eta_0\lambda_{\min}^K\sqrt{n\log(2n)}\sqrt{\xi(m)} =: \Delta(m) \end{aligned} \tag{5.180}$$

Summing the bound (5.170), we conclude that

$$\begin{aligned}&\|F(t)-F^{\mathrm{lin}}(t)\|_2 \\ &\leq \frac{9R_0^2}{\lambda_{\min}^K}\eta_0\sqrt{n}\log(2n)\left(\eta_0^2\frac{\pi^2c_0}{3}\left(2\lambda_{\min}^K\right)^2\frac{N(m)|\mathcal{N}|}{Lm}+8\,n^2\frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m)N^3(m)}\right) \\ &+\frac{3R_0R_2}{(\lambda_{\min}^K)^2}n^2\sqrt{n}\log(2n)\frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m)N^3(m)}+\Delta(m) \\ &\leq C_0\sqrt{n}\log(2n)\left(\eta_0^3\lambda_{\min}^K\frac{N(m)|\mathcal{N}|}{Lm}+n^2\left(\frac{\eta_0}{\lambda_{\min}^K}+\frac{1}{(\lambda_{\min}^K)^2}\right)\frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m)N^3(m)}\right) \\ &+\Delta(m) \end{aligned} \tag{5.181}$$

for some C_0. This can be simplified further by recalling Lemma 3.8, which ensures that for some $c>0$

$$m^2 \geq \frac{N^4(m)}{c|\mathcal{M}|^2|\mathcal{N}|^2} \qquad \to \qquad \frac{N(m)|\mathcal{N}|}{Lm} \leq \frac{cm|\mathcal{M}|^2|\mathcal{N}|^3}{LN^3(m)}, \tag{5.182}$$

and by noticing that, by assumption,

$$\eta_0 < \frac{2}{\lambda_{\min}^K+\lambda_{\max}^K} \leq \frac{1}{\lambda_{\min}^K}, \tag{5.183}$$

whence

$$\begin{aligned}&\|F(t)-F^{\mathrm{lin}}(t)\|_2 \\ &\leq C_1\sqrt{n}\log(2n)\left(\eta_0^3\lambda_{\min}^K+n^2\left(\frac{\eta_0}{\lambda_{\min}^K}+\frac{1}{(\lambda_{\min}^K)^2}\right)\right)\frac{Lm|\mathcal{M}|^4|\mathcal{N}|^3}{N^3(m)}+\Delta(m) \\ &\leq C_1\sqrt{n}\log(2n)\frac{1+2n^2}{(\lambda_{\min}^K)^2}\frac{Lm|\mathcal{M}|^4|\mathcal{N}|^3}{N^3(m)}+\Delta(m) \end{aligned}$$

$$\leq \frac{C}{(\lambda^K_{\min})^2} n^2 \sqrt{n} \log(2n) \frac{Lm|\mathcal{M}|^4|\mathcal{N}|^3}{N^3(m)} + \Delta(m) \tag{5.184}$$

for some C_1 and C. □

Now we are ready to prove Lemma 5.6. Many estimates are similar to the continuous time setting. Let us start from

$$\begin{aligned}
&\|\Delta(\Theta_t - \Theta^{\text{lin}}_t)\|_\infty \\
&\quad \leq \frac{\eta}{n} \left\| \nabla_\Theta f(\Theta_t, X^T)(F(t) - Y) + n\epsilon(t) - \nabla_\Theta f(\Theta_0, X^T)(F^{\text{lin}}(t) - Y) \right\|_\infty \\
&\quad \leq \frac{\eta}{n} \sup_i \|\partial_{\theta_i} f(\Theta_t, X) - \partial_{\theta_i} f(\Theta_0, X)\|_2 \|F(t) - Y\|_2 \\
&\qquad + \frac{\eta}{n} \sup_i \|\partial_{\theta_i} f(\Theta_0, X)\|_2 \|F^{\text{lin}}(t) - F(t)\|_2 + \eta \|\epsilon(t)\|_\infty.
\end{aligned} \tag{5.185}$$

Similarly to the continuous time setting,

$$\begin{aligned}
&\frac{\eta}{n} \sup_i \|\partial_{\theta_i} f(\Theta_t, X) - \partial_{\theta_i} f(\Theta_0, X)\|_2 \|F(t) - Y\|_2 \\
&\qquad \leq \frac{4\eta_0 R_0 R_1}{\lambda^K_{\min}} n^2 \log(2n) \frac{|\mathcal{M}|^3|\mathcal{N}|}{N^2_K(m) N^2(m)} \left(1 - \frac{\eta_0 \lambda^K_{\min}}{3}\right)^t =: A(t),
\end{aligned} \tag{5.186}$$

$$\begin{aligned}
&\frac{\eta}{n} \sup_i \|\partial_{\theta_i} f(\Theta_0, X)\|_2 \|F^{\text{lin}}(t) - F(t)\|_2 \\
&\qquad \leq 4R_0 \eta_0 \sqrt{\log(2n)} \frac{|\mathcal{M}|}{N(m)} \left(1 - \frac{\eta_0 \lambda^K_{\min}}{3}\right)^t =: C(t),
\end{aligned} \tag{5.187}$$

while, using Lemma 5.15,

$$\begin{aligned}
&\frac{\eta}{n} \sup_i \|\partial_{\theta_i} f(\Theta_0, X)\|_2 \|F^{\text{lin}}(t) - F(t)\|_2 \\
&\quad \leq \frac{\eta}{n} 2\sqrt{n} \frac{|\mathcal{M}|}{N(m)} \|F^{\text{lin}}(t) - F(t)\|_2 \\
&\quad \leq \frac{\eta}{n} 2\sqrt{n} \frac{|\mathcal{M}|}{N(m)} \left(\frac{C}{(\lambda^K_{\min})^2} n^2 \sqrt{n} \log(2n) \frac{Lm|\mathcal{M}|^4|\mathcal{N}|^3}{N^3(m)} + \Delta(m) \right) \\
&\quad = \frac{2\eta_0 C}{(\lambda^K_{\min})^2} n^3 \log(2n) \frac{Lm|\mathcal{M}|^5|\mathcal{N}|^3}{N_K(m) N^4(m)} + 2\eta_0 \sqrt{n} \frac{|\mathcal{M}|}{N_K(m) N(m)} \Delta(m) := B(t)
\end{aligned} \tag{5.188}$$

Therefore, (5.185) implies the following two bounds:

$$\|\Delta(\Theta_t - \Theta^{\text{lin}}_t)\|_\infty \leq A(t) + B(t) + \eta \|\epsilon(t)\|_\infty \tag{5.189}$$

and

$$\|\Delta(\Theta_t - \Theta^{\text{lin}}_t)\|_\infty \leq A(t) + C(t) + \eta \|\epsilon(t)\|_\infty. \tag{5.190}$$

Due to the hypothesis (5.41) the variance of $g^{(t)}(\Theta_t)$ is small enough for the bounds we are interested in: indeed, with probability at least $1 - \frac{\delta}{4}$, by (5.178) we have

$$\begin{aligned}
\|\epsilon(t)\|_\infty \leq \pi \sqrt{\frac{|\Theta|}{3\delta}} (t+1) \Bigg(& c_0 \eta_0^2 \frac{\left(\lambda^K_{\min}\right)^4}{n^2} \frac{N^2_K(m) N^2(m)}{|\mathcal{M}|^2 |\Theta|^3} \times \\
& \times \frac{\delta/4}{(t+1)^2} \frac{R_0^2}{2} \log(2n) \left(1 - \frac{1}{3} \eta_0 \lambda^K_{\min}\right)^{2t} \Bigg)^{1/2} \sqrt{\xi(m)},
\end{aligned}$$

$$= \frac{R_0\pi}{2}\sqrt{\log(2n)}\frac{\eta_0}{n}(\lambda_{\min}^K)^2\sqrt{\frac{c_0}{6}}\frac{N_K(m)N(m)}{|\mathcal{M}|Lm}\left(1-\frac{1}{3}\eta_0\lambda_{\min}^K\right)^t\sqrt{\xi(m)} \tag{5.191}$$

whence, similarly to (5.180)

$$\begin{aligned}\sum_{t=0}^{\infty}\eta\|\epsilon(t)\|_\infty &\le \frac{\pi R_0/2}{\frac{1}{3}\eta_0\lambda_{\min}^K}\sqrt{\frac{c_0}{6}}\sqrt{\log(2n)}\left(\eta_0\lambda_{\min}^K\right)^2\frac{N(m)}{|\mathcal{M}|Lm}\sqrt{\xi(m)}\\ &= \frac{1}{6\sqrt{n}}\left(9\pi R_0\sqrt{\frac{c_0}{6}}\sqrt{n\log(2n)}\eta_0\lambda_{\min}^K\sqrt{\xi(m)}\right)\frac{N(m)}{|\mathcal{M}|Lm}\\ &= \frac{1}{6\sqrt{n}}\frac{N(m)}{|\mathcal{M}|Lm}\Delta(m).\end{aligned} \tag{5.192}$$

As in the continuous time setting, we define

$$t^* = \left\lfloor \frac{3}{\eta_0\lambda_{\min}^K}\log N(m)\right\rfloor. \tag{5.193}$$

It is useful to notice that

$$1+x\le e^x \quad \forall\, x\in\mathbb{R} \qquad \to \qquad (1-x)^t \le e^{-xt} \quad \forall\, x\le 1 \quad \forall\, t\ge 0. \tag{5.194}$$

Indeed,

$$\begin{aligned}\sum_{t=t^*}^{\infty}\left(1-\frac{\eta_0\lambda_{\min}^K}{3}\right)^t &= \frac{3}{\eta_0\lambda_{\min}^K} - \sum_{t=0}^{t^*-1}\left(1-\frac{\eta_0\lambda_{\min}^K}{3}\right)^t\\ &= \frac{3}{\eta_0\lambda_{\min}^K} - \frac{1-(1-\eta_0\lambda_{\min}^K/3)^{t^*}}{\eta_0\lambda_{\min}^K/3}\\ &= \frac{3}{\eta_0\lambda_{\min}^K}\left(1-\frac{\eta_0\lambda_{\min}^K}{3}\right)^{t^*}\\ &\le \frac{3}{\eta_0\lambda_{\min}^K}e^{-\frac{1}{3}\eta_0\lambda_{\min}^K t^*} \le \frac{3}{\eta_0\lambda_{\min}^K}e^{-\log N(m)} = \frac{3}{\eta_0\lambda_{\min}^K}\frac{1}{N(m)}\end{aligned} \tag{5.195}$$

So we can sum the bounds for (5.185), using (5.195) for the one concerning $C(t)$:

$$\begin{aligned}\left\|\Theta_t-\Theta_t^{\text{lin}}\right\|_\infty &\le \sum_{t=0}^{\infty}A(t)+\sum_{t=0}^{t^*-1}B(t)+\sum_{t=t^*}^{\infty}C(t)+\sum_{t=0}^{\infty}\eta\|\epsilon(t)\|_\infty\\ &\le \frac{12R_0R_1}{(\lambda_{\min}^K)^2}n^2\log(2n)\frac{|\mathcal{M}|^3|\mathcal{N}|}{N_K^2(m)N^2(m)}\\ &\quad+\frac{6C}{(\lambda_{\min}^K)^3}n^3\log(2n)\frac{Lm|\mathcal{M}|^5|\mathcal{N}|^3}{N_K(m)N^4(m)}\log N(m)\\ &\quad+\frac{6}{\lambda_{\min}^K}\sqrt{n}\frac{|\mathcal{M}|}{N_K(m)N(m)}\log N(m)\Delta(m) \qquad (5.196)\\ &\quad+\frac{12R_0}{\lambda_{\min}^K}\sqrt{\log(2n)}\frac{|\mathcal{M}|}{N^2(m)}\\ &\quad+\frac{3}{2\sqrt{n}}\frac{N(m)}{|\mathcal{M}|Lm}\Delta(m). \qquad (5.197)\end{aligned}$$

Now, using the definition (5.145) of $\Delta(m)$ and recalling that $\eta_0\lambda_{\min}^K<1$

$$\Delta(m) = C'\eta_0\lambda_{\min}^K\sqrt{n\log(2n)}\sqrt{\xi(m)} < C'\sqrt{n\log(2n)}\sqrt{\xi(m)} \tag{5.198}$$

we conclude that, thanks to the estimate (5.182),

$$\begin{aligned}\left\|\Theta_t - \Theta_t^{\text{lin}}\right\|_\infty \leq{}& \frac{12R_0R_1\sqrt{c}}{(\lambda_{\min}^K)^2}n^2\log(2n)\frac{|\mathcal{M}|^4|\mathcal{N}|^2}{N_K^2(m)N^4(m)}\\ &+\frac{6C}{(\lambda_{\min}^K)^3}n^3\log(2n)\frac{Lm|\mathcal{M}|^5|\mathcal{N}|^3}{N_K(m)N^4(m)}\log N(m)\\ &+\frac{6C'\sqrt{c}}{\lambda_{\min}^K}n\sqrt{\log(2n)}\frac{Lm|\mathcal{M}|^2|\mathcal{N}|}{N_K(m)N^3(m)}\log N(m)\sqrt{\xi(m)}\\ &+\frac{12R_0\sqrt{c}}{\lambda_{\min}^K}\sqrt{\log(2n)}\frac{Lm|\mathcal{M}|^2|\mathcal{N}|}{N^4(m)}\\ &+\frac{3C'c}{2}\sqrt{\log(2n)}\frac{Lm|\mathcal{M}||\mathcal{N}|^2}{N^3(m)}\sqrt{\xi(m)}.\end{aligned} \tag{5.199}$$

Since $\lim_{m\to\infty} N(m) = \infty$ under our assumptions, we can suppose to take m large enough so that $\log N(m) \geq 1$. Therefore, we can finally simplify the above expression as follows:

$$\begin{aligned}&\left\|\Theta_t - \Theta_t^{\text{lin}}\right\|_\infty\\ &\quad\leq\left(\frac{C_1}{(\lambda_{\min}^K)^3}+C_2\right)n^3\log(2n)\frac{Lm|\mathcal{M}|^5|\mathcal{N}|^3}{N^4(m)}\log N(m)\frac{1}{2}\left(1+N(m)\sqrt{\xi(m)}\right).\end{aligned} \tag{5.200}$$

5.7 Proof of Theorem 5.8

Let $r \in \{0,1\}$ be the binary variable indicating the presence ($r=1$) or the absence ($r=0$) of noise. We will use it to unify (5.11) with (5.22):

$$\begin{aligned}|f(\Theta_t,x) - f^{\text{lin}}(\Theta_t,x)| \leq{}& \frac{M_1}{(\lambda_{\min}^K)^2}n^2\log(2n)\frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m)N^3(m)}\\ &+rM_2(\lambda_{\min}^K)^2\eta_0^2\log(2n)\frac{|\mathcal{N}|N(m)}{Lm}.\end{aligned} \tag{5.201}$$

Therefore, considering also (4.117) and (5.182)

$$\begin{aligned}&|f(\Theta_t,x) - f^{\text{lin}}(\Theta_t^{\text{lin}},x)|\\ &\quad= |f(\Theta_t,x) - f(\Theta_0,x) - \nabla_\Theta f(\Theta_0,x)^T(\Theta_t^{\text{lin}} - \Theta_0)|\\ &\quad= |f(\Theta_t,x) - f(\Theta_0,x) - \nabla_\Theta f(\Theta_0,x)^T(\Theta_t - \Theta_0) - \nabla_\Theta f(\Theta_0,x)^T(\Theta_t^{\text{lin}} - \Theta_t)|\\ &\quad\leq |f(\Theta_t,x) - f^{\text{lin}}(\Theta_t,x)| + |\nabla_\Theta f(\Theta_0,x)^T(\Theta_t^{\text{lin}} - \Theta_t)|\\ &\quad\leq |f(\Theta_t,x) - f^{\text{lin}}(\Theta_t,x)| + \|\nabla_\Theta f(\Theta_0,x)\|_1\|\Theta_t^{\text{lin}} - \Theta_t\|_\infty\\ &\quad\leq \frac{M_1}{(\lambda_{\min}^K)^2}n^2\log(2n)\frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m)N^3(m)} + rM_2c(\lambda_{\min}^K)^2\eta_0^2\log(2n)\frac{Lm|\mathcal{M}|^2|\mathcal{N}|^3}{N^3(m)}\\ &\qquad+2\frac{Lm|\mathcal{M}|}{N(m)}\|\Theta_t^{\text{lin}} - \Theta_t\|_\infty.\end{aligned} \tag{5.202}$$

Combining this result with Corollary (5.7), we have that

$$\begin{aligned}&|f(\Theta_t,x) - f^{\text{lin}}(\Theta_t^{\text{lin}},x)|\\ &\quad\leq \frac{M_1}{(\lambda_{\min}^K)^2}n^2\log(2n)\frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N_K^2(m)N^3(m)} + rM_2c(\lambda_{\min}^K)^2\eta_0^2\log(2n)\frac{Lm|\mathcal{M}|^2|\mathcal{N}|^3}{N^3(m)}\\ &\qquad+2\left(\frac{C_1}{(\lambda_{\min}^K)^3}+C_2\right)n^3\log(2n)\frac{L^2m^2|\mathcal{M}|^6|\mathcal{N}|^3}{N^5(m)}\log N(m)\\ &\quad\leq \frac{M_1}{(\lambda_{\min}^K)^2}n^2\log(2n)\frac{L^2m^2|\mathcal{M}|^5|\mathcal{N}|^2}{N^5(m)} + rM_2c\log(2n)\frac{L^2m^2|\mathcal{M}|^3|\mathcal{N}|^4}{N^3(m)}\end{aligned}$$

$$+2\left(\frac{C_1}{(\lambda^K_{\min})^3}+C_2\right)n^3\log(2n)\frac{L^2m^2|\mathcal{M}|^6|\mathcal{N}|^3}{N^5(m)}\log N(m)$$
$$\leq\left(\frac{C_3}{(\lambda^K_{\min})^3}+C_4\right)n^3\log(2n)\frac{L^2m^2|\mathcal{M}|^6|\mathcal{N}|^4}{N^5(m)}\log N(m), \tag{5.203}$$

where we supposed, as above, that m is large enough so that $\log N(m)\geq 1$ and, in order to get the above bound simplified, we also used (5.183) and Lemma 3.8 in order to get an inequality as in (5.182).

5.8 Proof of Theorem 5.9

Lemma 5.16 *The hypothesis*

$$\lim_{m\to\infty}\frac{L^2m^2|\mathcal{M}|^6|\mathcal{N}|^4}{N^5(m)}\log N(m)=0, \tag{5.204}$$

of Theorem 5.9 ensures that the hypotheses of Theorem 3.7, of Theorem 4.7 and of Theorem 5.4 are satisfied, i.e.

$$\lim_{m\to\infty}\frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3(m)}=0, \tag{5.205}$$

$$\lim_{m\to\infty}\frac{1}{N^2_K(m)}\frac{\Sigma_2|\mathcal{M}|^2|\mathcal{N}|^2}{N^4(m)}=0, \tag{5.206}$$

$$\lim_{m\to\infty}\frac{Lm|\mathcal{M}|^4|\mathcal{N}|^2}{N^3(m)}=0. \tag{5.207}$$

Furthermore, the bounds (5.21) and (5.22) of Theorem 5.4 are nontrivial as $m\to\infty$, i.e.

$$\lim_{m\to\infty}\frac{N(m)}{|\mathcal{M}||\Theta|}=0,\qquad \lim_{m\to\infty}\frac{|\mathcal{M}|}{N_K(m)N(m)}=0, \tag{5.208}$$

$$\lim_{m\to\infty}\frac{|\mathcal{N}|N(m)}{Lm}=0,\qquad \lim_{m\to\infty}\frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N^2_K(m)N^3(m)}=0. \tag{5.209}$$

Proof (5.205) and (5.206) are ensured by (5.207) because of Lemma 5.12. So we just need to prove (5.207). Using the inequality of Lemma 3.8 in the form

$$1\leq c\frac{m|\mathcal{M}||\mathcal{N}|}{N^2(m)} \tag{5.210}$$

we have

$$0\leq\lim_{m\to\infty}\frac{Lm|\mathcal{M}|^4|\mathcal{N}|^2}{N^3(m)}\leq\lim_{m\to\infty}\frac{Lm^2|\mathcal{M}|^5|\mathcal{N}|^3}{N^5(m)}$$
$$\leq\lim_{m\to\infty}\frac{L^2m^2|\mathcal{M}|^6|\mathcal{N}|^4}{N^5(m)}\log N(m)=0 \tag{5.211}$$

Now we notice that

$$0\leq\lim_{m\to\infty}\frac{N(m)}{|\mathcal{M}||\Theta|}\leq\lim_{m\to\infty}\frac{|\mathcal{N}|N(m)}{Lm}=\lim_{m\to\infty}\frac{|\mathcal{N}|N^6(m)}{LmN^5(m)}$$
$$\leq c^3\lim_{m\to\infty}\frac{|\mathcal{N}|(m^3|\mathcal{M}|^3|\mathcal{N}|^3)}{LmN^5(m)}\leq c^3\lim_{m\to\infty}\frac{L^2m^2|\mathcal{M}|^6|\mathcal{N}|^4}{N^5(m)}\log N(m)=0, \tag{5.212}$$

where we used again 5.210. This proves the first requirements of (5.208) and (5.209). Finally, with a similar strategy

$$0\leq\lim_{m\to\infty}\frac{|\mathcal{M}|}{N_K(m)N(m)}\leq c^2\lim_{m\to\infty}\frac{m^2|\mathcal{M}|^3|\mathcal{N}|^2}{N^5(m)}$$
$$\leq c^2\lim_{m\to\infty}\frac{L^2m^2|\mathcal{M}|^6|\mathcal{N}|^4}{N^5(m)}\log N(m)=0, \tag{5.213}$$

and

$$0\leq\lim_{m\to\infty}\frac{Lm|\mathcal{M}|^4|\mathcal{N}|}{N^2_K(m)N^3(m)}\leq\lim_{m\to\infty}c\frac{Lm^2|\mathcal{M}|^5|\mathcal{N}|^4}{N^5(m)}$$
$$\leq\lim_{m\to\infty}c\frac{L^2m^2|\mathcal{M}|^6|\mathcal{N}|^4}{N^5(m)}\log N(m)=0. \tag{5.214}$$

This concludes the proof of Lemma 5.16. □

As in section 4, we need to show that the linearized model converges to a Gaussian process in the limit of many qubits. It will be useful to recall that

$$\sum_{k=0}^{t-1}\left(\mathbb{1}-\eta_0\hat{K}_{\Theta_0}\right)^k=\frac{\mathbb{1}-\left(\mathbb{1}-\eta_0\hat{K}_{\Theta_0}\right)^t}{\mathbb{1}-\left(\mathbb{1}-\eta_0\hat{K}_{\Theta_0}\right)}=\frac{1}{\eta_0}\left(\mathbb{1}-\left(\mathbb{1}-\eta_0\hat{K}_{\Theta_0}\right)^t\right)\hat{K}_{\Theta_0}^{-1}, \tag{5.215}$$

provided that $\hat{K}_{\Theta_0}$ is invertible.

Lemma 5.17 *The solution of the gradient descent equation (5.3) for the linearized model is*

$$f^{\text{lin}}(\Theta_t,x)=f(\Theta_0,x)-\hat{K}_{\Theta_0}(x,X^T)\left(\mathbb{1}-\left(\mathbb{1}-\eta_0\hat{K}_{\Theta_0}\right)^t\right)\hat{K}_{\Theta_0}^{-1}(f(\Theta_0,X)-Y), \tag{5.216}$$

provided that $\hat{K}_{\Theta_0}$ is invertible.

Proof We start from (5.3) for the linearized model:

$$\Theta_{t+1}-\Theta_t=-\frac{\eta}{n}\nabla_\Theta f^{\text{lin}}(\Theta_t,X^T)(f^{\text{lin}}(\Theta_t,X)-Y), \tag{5.217}$$

which can be rewritten as

$$\begin{aligned}\bar{\Theta}_{t+1}-\bar{\Theta}_t&=-\frac{\eta}{n}\nabla_\Theta f(\Theta_0,X^T)(f(\Theta_0,X)+\nabla_\Theta^T f(\Theta_0,X)(\bar{\Theta}_t-Y)\\&=-\frac{\eta}{n}\nabla_\Theta f(\Theta_0,X^T)\nabla_\Theta^T f(\Theta_0,X)\,\bar{\Theta}_t+b,\end{aligned} \tag{5.218}$$

where $\bar{\Theta}_t\equiv\Theta_t-\Theta_0$ and

$$b=-\frac{\eta}{n}\nabla_\Theta f(\Theta_0,X^T)(f(\Theta_0,X)-Y). \tag{5.219}$$

So, we have to solve the recurrence relation

$$\begin{cases}\bar{\Theta}_{t+1}=\left(\mathbb{1}-\frac{\eta}{n}\nabla_\Theta f(\Theta_0,X^T)\nabla_\Theta^T f(\Theta_0,X)\right)\bar{\Theta}_t+b\\ \bar{\Theta}_0=0\end{cases}. \tag{5.220}$$

Calling

$$M_{\Theta_0}=\frac{1}{N_K(m)}\nabla_\Theta f(\Theta_0,X^T)\nabla_\Theta^T f(\Theta_0,X) \tag{5.221}$$

and redefining $\eta=\frac{n}{N_K(m)}\eta_0$, we have

$$\begin{cases}\bar{\Theta}_{t+1}=(\mathbb{1}-\eta_0 M_{\Theta_0})\,\bar{\Theta}_t+b\\ \bar{\Theta}_0=0\end{cases}. \tag{5.222}$$

It is easy to see that

$$\begin{aligned}\bar{\Theta}_1&=b\\ \bar{\Theta}_2&=b+(\mathbb{1}-\eta_0 M_{\Theta_0})\,b\\ \bar{\Theta}_3&=b+(\mathbb{1}-\eta_0 M_{\Theta_0})\,b+(\mathbb{1}-\eta_0 M_{\Theta_0})^2\,b\\ &\vdots\\ \bar{\Theta}_t&=\sum_{k=0}^{t-1}(\mathbb{1}-\eta_0 M_{\Theta_0})^k\,b,\end{aligned} \tag{5.223}$$

whence

$$\begin{aligned}f^{\text{lin}}(\Theta_t,x)&=f(\Theta_0,x)+\nabla_\Theta^T f(\Theta_0,x)\bar{\Theta}_t\\&=f(\Theta_0,x)+\nabla_\Theta^T f(\Theta_0,x)\sum_{k=0}^{t-1}(\mathbb{1}-\eta_0 M_{\Theta_0})^k\,b\\&=f(\Theta_0,x)-\frac{\eta}{n}\nabla_\Theta^T f(\Theta_0,x)\sum_{k=0}^{t-1}(\mathbb{1}-\eta_0 M_{\Theta_0})^k\,\nabla_\Theta f(\Theta_0,X^T)(f(\Theta_0,X)-Y).\end{aligned} \tag{5.224}$$

By the definition of M_{Θ_0}.

$$
\begin{aligned}
&M_{\Theta_0}\nabla_\Theta f(\Theta_0, X^T) \\
&\quad = \frac{1}{N_K(m)}\Big(\nabla_\Theta f(\Theta_0, X^T)\nabla_\Theta^T f(\Theta_0, X)\Big)\nabla_\Theta f(\Theta_0, X^T) \\
&\quad = \nabla_\Theta f(\Theta_0, X^T)\frac{1}{N_K(m)}\Big(\nabla_\Theta^T f(\Theta_0, X)\nabla_\Theta f(\Theta_0, X^T)\Big) \\
&\quad = \nabla_\Theta f(\Theta_0, X^T)\hat{K}_{\Theta_0}, \qquad (5.225)
\end{aligned}
$$

which yields

$$
\begin{aligned}
f^{\text{lin}}(\Theta_t, x) &= f(\Theta_0, x) - \frac{\eta}{n}\nabla_\Theta^T f(\Theta_0, x)\nabla_\Theta f(\Theta_0, X^T)\sum_{k=0}^{t-1}\left(\mathbb{1} - \eta_0\hat{K}_{\Theta_0}\right)^k (f(\Theta_0, X) - Y) \\
&= f(\Theta_0, x) - \eta_0\hat{K}_{\Theta_0}(x, X^T)\sum_{k=0}^{t-1}\left(\mathbb{1} - \eta_0\hat{K}_{\Theta_0}\right)^k (f(\Theta_0, X) - Y) \\
&= f(\Theta_0, x) - \hat{K}_{\Theta_0}(x, X^T)\left(\mathbb{1} - \left(\mathbb{1} - \eta_0\hat{K}_{\Theta_0}\right)^t\right)\hat{K}_{\Theta_0}^{-1}(f(\Theta_0, X) - Y). \qquad (5.226)
\end{aligned}
$$

□

Remark 5.4 Let

$$
\eta_0 < \frac{2}{\lambda_{\min}^K + \lambda_{\max}^K}, \qquad (5.227)
$$

i.e.

$$
\eta_0 = \frac{2}{\lambda_{\min}^K + \lambda_{\max}^K + 2\epsilon} \qquad \text{for some} \qquad \epsilon > 0. \qquad (5.228)
$$

Because of Assumption 4.3 and Lemma 4.25, $\hat{K}_{\Theta_0}(X, X^T)$ converges to $\bar{K}$, which is invertible, as $m \to \infty$. So, if m is large enough, then

$$
\text{Spec}(\hat{K}_{\Theta_0}) \subseteq \left[\frac{\lambda_{\min}^K}{2}, \lambda_{\max}^K + \epsilon\right], \qquad (5.229)
$$

whence

$$
1 - \eta_0\frac{\lambda_{\min}^K}{2} < 1, \qquad (5.230)
$$

$$
1 - \eta_0(\lambda_{\max}^K + \epsilon) > 1 - \frac{2\lambda_{\max} + 2\epsilon}{\lambda_{\max}^K + 2\epsilon} = -1 + \frac{2\epsilon}{\lambda_{\max}^K + 2\epsilon} > -1. \qquad (5.231)
$$

Therefore

$$
-1 < 1 - \eta_0\lambda < 1 \qquad \forall\, \lambda \in \text{Spec}(\hat{K}_{\Theta_0}) \qquad \rightarrow \qquad -\mathbb{1} \prec \mathbb{1} - \eta_0\hat{K}_{\Theta_0} \prec \mathbb{1} \qquad (5.232)
$$

so, in the limit $t \to \infty$,

$$
\left(\mathbb{1} - \eta_0\hat{K}_{\Theta_0}\right)^t \to 0 \qquad (5.233)
$$

and (5.226) converges to

$$
\lim_{t\to\infty} f^{\text{lin}}(\Theta_t, x) = f(\Theta_0, x) - \hat{K}_{\Theta_0}(x, X^T)\hat{K}_{\Theta_0}^{-1}(f(\Theta_0, X) - Y). \qquad (5.234)
$$

Lemma 5.18 *In the limit $m \to \infty$, $\{f^{\text{lin}}(\Theta_t^{\text{lin}}, x)\}_{x\in\mathcal{X}}$ converges in distribution to a Gaussian process $\{f_t^{(\infty)}(x)\}_{x\in\mathcal{X}}$ with mean and variance*

$$
\begin{aligned}
\mu_t(x) &= \bar{K}(x, X^T)K^{-1}\left(\mathbb{1} - \left(\mathbb{1} - \eta_0\bar{K}\right)^t\right)\bar{Y}, \qquad (5.235) \\
\mathcal{K}_t(x, x') &= \mathcal{K}_0(x, x'), \\
&\quad - \bar{K}(x, X^T)\bar{K}^{-1}\left(\mathbb{1} - \left(\mathbb{1} - \eta_0\bar{K}\right)^t\right)\mathcal{K}_0(X, x') \\
&\quad - \bar{K}(x', X^T)\bar{K}^{-1}\left(\mathbb{1} - \left(\mathbb{1} - \eta_0\bar{K}\right)^t\right)\mathcal{K}_0(X, x) \\
&\quad + \bar{K}(x, X^T)\bar{K}^{-1}\left(\mathbb{1} - \left(\mathbb{1} - \eta_0\bar{K}\right)^t\right)\times \\
&\qquad \times \mathcal{K}_0(X, X^T)\left(\mathbb{1} - \left(\mathbb{1} - \eta_0\bar{K}\right)^t\right)\bar{K}^{-1}\bar{K}(X, x'). \qquad (5.236)
\end{aligned}
$$

Proof By Lemma 4.25,

$$\hat{K}_\Theta(x,x') \xrightarrow{p} \bar{K}(x,x') \quad \text{as} \quad m\to\infty. \tag{5.237}$$

Furthermore, by Theorem 3.7

$$f(\Theta_0,\,\cdot\,) \xrightarrow{d} f^{(\infty)}(\,\cdot\,) \quad \text{as} \quad m\to\infty. \tag{5.238}$$

Let $\mathcal{F}=\{x_\alpha\}_{\alpha\in A}$ be a finite family of inputs $x_\alpha\in\mathcal{X}$ containing the inputs of the dataset

$$\{x^{(i)}\}_{1\le i\le n}\subseteq\mathcal{F} \tag{5.239}$$

Since the solution (5.226) for the output corresponding to any input $x_\beta\in\mathcal{F}_A$ is the following linear combination of the outputs $\{f(\Theta_0,x_\alpha)\}_{\alpha\in A}$ and of Y,

$$f^{\text{lin}}(\Theta_t^{\text{lin}},x_\beta)=f(\Theta_0,x_\beta)-\eta_0\hat{K}_{\Theta_0}(x,X^T)\left(\mathbb{1}-\left(\mathbb{1}-\eta_0\hat{K}_{\Theta_0}\right)^t\right)\hat{K}_{\Theta_0}^{-1}(f(\Theta_0,X)-Y), \tag{5.240}$$

we can write

$$f^{\text{lin}}(\Theta_t^{\text{lin}},x_\beta)=\sum_{\alpha\in A}\tilde{M}^{(t)}_{\beta\alpha}[\hat{K}_{\Theta_0}]f(\Theta_0,x_\alpha)+\left(\tilde{R}^{(t)}[\hat{K}_{\Theta_0}]\right)^T Y, \tag{5.241}$$

where the entries $\tilde{M}^{(t)}_{\beta\alpha}[\hat{K}_{\Theta_0}]$ and the components of $\tilde{R}^{(t)}[\hat{K}_{\Theta_0}]$ are continous functions of the elements of matrix of the empirical NTK

$$\{\hat{K}_{\Theta_0}(x_\alpha,x_{\alpha'})\}_{\alpha,\alpha'\in A}. \tag{5.242}$$

By continuity, the (finite) matrix $\tilde{M}^{(t)}_{\beta\alpha}[\hat{K}_{\Theta_0}]$ and the (finite) vector $\tilde{R}^{(t)}[\hat{K}_{\Theta_0}]$ converge in probability to $\tilde{M}^{(t)}_{\beta\alpha}[\bar{K}]$ and $\tilde{R}^{(t)}[\bar{K}]$:

$$\tilde{M}^{(t)}_{\beta\alpha}[\hat{K}_{\Theta_0}]\xrightarrow{p}\tilde{M}^{(t)}_{\beta\alpha}[\bar{K}],\qquad\qquad \tilde{R}^{(t)}[\hat{K}_{\Theta_0}]\xrightarrow{p}\tilde{R}^{(t)}[\bar{K}]. \tag{5.243}$$

By Slutsky's theorem 4.32, we conclude that

$$\{f^{\text{lin}}(\Theta_t^{\text{lin}},x_\beta)\}_{x_\beta\in\mathcal{F}}\xrightarrow{d}\left\{\sum_{\alpha\in A}\tilde{M}^{(t)}_{\beta\alpha}[\bar{K}]f^{(\infty)}(x_\alpha)+\left(R^{(t)}[\bar{K}]\right)^T Y\right\}_{x_\beta\in\mathcal{F}}\quad\text{as}\quad m\to\infty, \tag{5.244}$$

i.e.

$$f^{\text{lin}}(\Theta_t^{\text{lin}},\,\cdot\,)\big|_{\mathcal{F}}\xrightarrow{d}f^{(\infty)}(\,\cdot\,)\big|_{\mathcal{F}}-\bar{K}(\,\cdot\,,X^T)\big|_{\mathcal{F}}\left(\mathbb{1}-\left(\mathbb{1}-\eta_0\bar{K}\right)^t\right)\bar{K}^{-1}(f^{(\infty)}(X)-Y). \tag{5.245}$$

In the limit $m\to\infty$, the solution (5.245) is a linear combination of the Gaussian processes $\{f^{\text{lin}}(\Theta_0,x)\}_{x\in\mathcal{F}}=\{f(\Theta_0,x)\}_{x\in\mathcal{F}}$ and $F(0)$, so it is a Gaussian process $\{f_t^{(\infty)}(x)\}_{x\in\mathcal{F}}$ as well, with

$$\begin{aligned}
\mu_t(x)&=\mathbb{E}\left[f_t^{(\infty)}(x)\right]\\
&=\mathbb{E}\left[f^{(\infty)}(x)\right]-\bar{K}(x,X^T)\left(\mathbb{1}-\left(\mathbb{1}-\eta_0\bar{K}\right)^t\right)\bar{K}^{-1}(\mathbb{E}\left[F^{(\infty)}\right]-Y)\\
&=\bar{K}(x,X^T)\left(\mathbb{1}-\left(\mathbb{1}-\eta_0\bar{K}\right)^t\right)\bar{K}^{-1}Y,
\end{aligned} \tag{5.246}$$

$$\begin{aligned}
\mathcal{K}_t(x,x')&=\mathbb{E}\left[\left(f_t^{(\infty)}(x)-\mu_t(x)\right)\left(f_t^{(\infty)}(x')-\mu_t(x')\right)\right]\\
&=\mathbb{E}\Bigg[\left(f^{(\infty)}(x)-\bar{K}(x,X^T)\left(\mathbb{1}-\left(\mathbb{1}-\eta_0\bar{K}\right)^t\right)\bar{K}^{-1}F^{(\infty)}\right)\times\\
&\qquad\times\left(f^{(\infty)}(x')-\bar{K}(x',X^T)\left(\mathbb{1}-\left(\mathbb{1}-\eta_0\bar{K}\right)^t\right)\bar{K}^{-1}F^{(\infty)}\right)\Bigg]\\
&=\mathcal{K}_0(x,x')\\
&\quad-\bar{K}(x,X^T)\bar{K}^{-1}\left(\mathbb{1}-\left(\mathbb{1}-\eta_0\bar{K}\right)^t\right)\mathcal{K}_0(X,x')\\
&\quad-\bar{K}(x',X^T)\bar{K}^{-1}\left(\mathbb{1}-\left(\mathbb{1}-\eta_0\bar{K}\right)^t\right)\mathcal{K}_0(X,x)\\
&\quad+\bar{K}(x,X^T)\bar{K}^{-1}\left(\mathbb{1}-\left(\mathbb{1}-\eta_0\bar{K}\right)^t\right)\bar{\times}\\
&\qquad\times\mathcal{K}_0(X,X^T)\left(\mathbb{1}-\left(\mathbb{1}-\eta_0\bar{K}\right)^t\right)\bar{K}^{-1}\bar{K}(X,x').
\end{aligned} \tag{5.247}$$

Since we assumed $\mathcal{X}$ to be finite, this is enough to prove the convergence of the distribution to the entire Gaussian process: it is sufficient to choose $\mathcal{F}=\mathcal{X}$. □

Remark 5.5 In section 6 we will generalize the convergence of $\{f(\Theta_t, x)\}_{x\in\mathcal{X}}$ to a Gaussian process for the case of $\mathcal{X}$ being infinite. This will not require to prove that also $\{f^{\text{lin}}(\Theta_t, x)\}_{x\in\mathcal{X}}$ converges to a Gaussian process: we will only need the convergence of $\{f(\Theta_t, x)\}_{x\in\mathcal{F}}$ for any $\mathcal{F}$ finite set of inputs, which, as we will see, is a corollary of the convergence of the linearized model for a finite number of inputs. Therefore, the proof given above will be enough for our purposes.

Now we are ready to prove Theorem 5.9.
Let $\bar{X} \in \mathbb{R}^N$ be any (finite dimensional) vector on N distinct inputs $\{\bar{x}_1, \dots, \bar{x}_n\} \in \mathcal{X}$ and let $\Delta_t(\bar{X})$ be the random vector

$$\Delta_t(\bar{X}) = f(\Theta_t, \bar{X}) - f^{\text{lin}}(\Theta_t, \bar{X}). \tag{5.248}$$

We show that $\Delta_t(\bar{X}) \xrightarrow{p} 0$ as $m \to \infty$. Let $\epsilon, \delta > 0$. By Theorem 5.8, there exists m_0 such that, for any $m \geq m_0$,

$$\mathbb{P}\Bigg(\sup_{t'} \sup_{x\in\mathcal{X}} |f(\Theta_{t'}, x) - f^{\text{lin}}(\Theta^{\text{lin}}_{t'}, x)| \\ \leq \left(\frac{C_3}{(\lambda^k_{\min})^3} + C_4 \right) n^3 \log(2n) \frac{L^2 m^2 |\mathcal{M}|^6 |\mathcal{N}|^4}{N^5(m)} \log N(m) \Bigg) \geq 1 - \delta. \tag{5.249}$$

whence

$$\mathbb{P}\Bigg(\|f(\Theta_t, \bar{X}) - f^{\text{lin}}(\Theta^{\text{lin}}_t, \bar{X})\|_2 \\ \leq \sqrt{N} \left(\frac{C_3}{(\lambda^k_{\min})^3} + C_4 \right) n^3 \log(2n) \frac{L^2 m^2 |\mathcal{M}|^6 |\mathcal{N}|^4}{N^5(m)} \log N(m) \Bigg) \geq 1 - \delta. \tag{5.250}$$

There is also $m_1 \in \mathbb{N}$ such that

$$\sqrt{N} \left(\frac{C_3}{(\lambda^k_{\min})^3} + C_4 \right) n^3 \log(2n) \frac{L^2 m^2 |\mathcal{M}|^6 |\mathcal{N}|^4}{N^5(m)} \log N(m) < \epsilon \qquad \forall\, m \geq m_1. \tag{5.251}$$

So, for any $m \geq \max\{m_0, m_1\}$,

$$\mathbb{P}\left(\|\Delta_t(\bar{X})\|_2 < \epsilon \right) \geq 1 - \delta, \tag{5.252}$$

so

$$\Delta_t(\bar{X}) \xrightarrow{p} 0 \quad \text{as} \quad m \to \infty. \tag{5.253}$$

By Corollary 4.9, $\{f^{\text{lin}}(\Theta^{\text{lin}}_t, x)\}_{x\in\mathcal{X}}$ converges in distribution to a Gaussian process with mean μ_t and covariance $\mathcal{K}_t$ defined in the statement of the corollary:

$$\{f^{\text{lin}}(\Theta_t, x)\}_{x\in\mathcal{X}} \xrightarrow{d} \{f^{(\infty)}_t(x)\}_{x\in\mathcal{X}}. \tag{5.254}$$

Now, applying Slutsky's theorem 4.32 we have

$$f(\Theta_t, \bar{X}) = f^{\text{lin}}(\Theta_t, \bar{X}) + \Delta_t(\bar{X}) \xrightarrow{d} f^{(\infty)}_t(x) \quad \text{as} \quad m \to \infty. \tag{5.255}$$

Since $\mathcal{X}$ is finite, we can choose $\bar{X} = \text{vec}(\mathcal{X})$ so that we have the convergence of the full multivariate Gaussian distribution, which is the Gaussian process itself. And this concludes the proof of Theorem 5.9.

Remark 5.6 This argument is not valid in the case of $\mathcal{X}$ infinite because we cannot consider $\bar{X} = \text{vec}(\mathcal{X})$. See section 6 for the proof in the most general case.

6 The case of an infinite input space

So far we considered the case of a finite input space $\mathcal{X}$ in order to provide simpler proofs for the reader interested in the physical case of having a finite number of digits to codify an input belonging to a compact space. However, for a rigorous and complete mathematical treatment of the topic, we will extend all our results to the more general case of an infinite input space. In particular, in order to prove the convergence when $\mathcal{X}$ is infinite, we need the following steps:

1. the functions generated by quantum neural networks are equicontinuous and equibounded with high probability (Lemma 6.2 and Lemma 6.3);
2. therefore, with high probability, they belong to a compact set by Ascoli-Arzelà's theorem;
3. as a consequence, Prokhorov's theorem can be invoked: the family of the distributions of the functions generated by quantum neural networks is compact;
4. we will use this fact to prove that the limit of that stochastic process is the Gaussian process defined by the linearized model (Theorem 6.4).

Let us see all the steps in detail. All the statements will be proved for the case of discrete time with noise for a more concise presentation. Most of the proofs in the continuous time setting would be identical or, when slightly different, simpler; the reader can easily reconstruct the complete proofs for that case.

In order to prove the equicontinuity of the model function (Lemma 6.2) we need the following concentration inequality:

Lemma 6.1 (Corollary of the Efron–Stein inequality [67]) *Let $X_1, \dots, X_n$ be independent random variables taking values in $\mathbb{X}$ and let $Z = f(X)$ be a square-integrable function of $X = (X_1, \dots, X_n)$ with the bounded differences property: for some constants $c_1, \dots, c_n$ it holds that*

$$\sup_{x_1,\dots,x_n,x_i' \in \mathbb{X}} \left| f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n) \right| \leq c_i \quad 1 \leq i \leq n. \tag{6.1}$$

Then,

$$\mathrm{Var}(Z) \leq \frac{1}{4} \sum_{i=1}^{n} c_i^2 \,. \tag{6.2}$$

Now we can prove the following lemma:

Lemma 6.2 (Equicontinuity with high probability) *Under the same hypotheses of Theorem 3.7, let us suppose that*

$$\lim_{m \to \infty} \frac{Lm|\mathcal{M}|^4 |\mathcal{N}|^2}{N_K^2(m) N^4(m)} = 0. \tag{6.3}$$

Then, for any $\delta > 0$ there exists $M > 0$ such that for any $m \geq 1$, with probability at least $1 - \delta$ we have

$$\max_{i=1,\dots,\dim \mathcal{X}} \sup_{\substack{x \in \mathcal{X} \\ t \geq 0}} |\partial_{x_i} f(\Theta_t, x)| \leq M \,. \tag{6.4}$$

Proof As introduced in subsubsection 2.2.2, $f(\Theta, x)$ is periodic with period π in each input x_i. So, we can consider the Fourier series decomposition

$$f(\Theta, x) = \sum_{q \in \mathbb{Z}^{\dim \mathcal{X}}} f^{(q)}(\Theta) e^{2iq \cdot x} \quad \text{where} \quad f^{(q)}(\Theta) = \int_{\mathcal{X}} \frac{dx}{(\pi)^{\dim \mathcal{X}}} e^{-2iq \cdot x} f(\Theta, x). \tag{6.5}$$

Recalling Definition 2.7 and Lemma 2.4, we can consider the pruning of the model (2.16):

$$\begin{aligned} f_k(\Theta, x) &= \frac{1}{N(m)} \left\langle 0 \middle| U^\dagger(\Theta, x) O_k U(\Theta, x) \middle| 0 \right\rangle \\ &= \frac{1}{N(m)} \left\langle 0 \middle| \left[U^\dagger(\Theta, x) \right]_k O_k \left[U(\Theta, x) \right]_k \middle| 0 \right\rangle \\ &= \frac{1}{N(m)} \langle 0 | \left[W_1^\dagger(\Theta) V_1^\dagger(x) \cdots W_L^\dagger(\Theta) V_L^\dagger(x) \right]_k \\ &\qquad O_k \left[V_L(x) W_L(\Theta) \cdots V_1(x) W_1(\Theta) \right]_k |0\rangle, \end{aligned} \tag{6.6}$$

where the only encoding unitaries in (2.15)

$$V_\ell(x) = \prod_{E\in S_\ell} \prod_{j=1}^{\dim \mathcal{X}} U_{E,j}^{(\ell)} \prod_{i\in E} e^{-ix_j \mathcal{K}_j^{\ell,i}} \tag{6.7}$$

which may contribute to $f_k(\Theta, x)$ are the ones such that $i \in \mathcal{N}_k^\ell$, as a consequence of the pruning property (see Lemma 2.4), and such that $\mathcal{K}_j^{\ell,i} \neq 0$. Therefore, let $N_{j,k}$ be the number of times that x_j appears nontrivially (i.e., $\mathcal{K}_j^{\ell,i} \neq 0$) in the light cone of $f_k(\Theta, x)$. We can consider the projectors $P^{\pm}_{j,\ell,i}$ on the eigenspaces of $\mathcal{K}_j^{\ell,i}$:

$$\mathcal{K}_j^{\ell,i} = P^+_{j,\ell,i} - P^-_{j,\ell,i} \quad P^+_{j,\ell,i} + P^-_{j,\ell,i} = \mathbb{1} \tag{6.8}$$

In this way, we can rewrite

$$\begin{aligned} e^{-ix_j \mathcal{K}_j^{\ell,i}} &= e^{-ix_j P^+_{j,\ell,i}} e^{ix_j P^-_{j,\ell,i}} \\ &= e^{-ix_j P^+_{j,\ell,i}} (P^+_{j,\ell,i} + P^-_{j,\ell,i}) e^{ix_j P^-_{j,\ell,i}} \\ &= e^{-ix_j} P^+_{j,\ell,i} + e^{ix_j} P^-_{j,\ell,i}. \end{aligned} \tag{6.9}$$

This means that in $f_k(\Theta, x)$ the dependence on x will appear as

$$f_k(\Theta, x) = \sum_{q\in Q_k} e^{2iq\cdot x} f^{(q,k)}(\Theta), \tag{6.10}$$

where $f^{(q,k)}(\Theta)$ does not depend on x and where

$$Q_k := \bigtimes_{j=1}^{\dim \mathcal{X}} \{-N_{j,k}, -N_{j,k}+1, \dots, N_{j,k}-1, N_{j,k}\}. \tag{6.11}$$

We define $N_{\max} = \max_{j,k} N_{j,k}$ so that we can extend

$$Q_k \subseteq Q := \{-N_{\max}, -N_{\max}+1, \dots, N_{\max}-1, N_{\max}\}^{\dim \mathcal{X}} \tag{6.12}$$

and

$$f^{(q,k)}(\Theta) := 0 \quad \text{when} \quad q \in Q \setminus Q_k, \tag{6.13}$$

so that we can rewrite

$$f_k(\Theta, x) = \sum_{q\in Q} e^{2iq\cdot x} f^{(q,k)}(\Theta). \tag{6.14}$$

Then

$$f(\Theta, x) = \sum_{k=1}^m f_k(\Theta, x) = \sum_{q\in Q} e^{2iq\cdot x_i} \left(\sum_{k=1}^m f^{(q,k)}(\Theta) \right). \tag{6.15}$$

Calling

$$f^{(q)}(\Theta) := \sum_{k=1}^m f^{(q,k)}(\Theta), \tag{6.16}$$

we have proved the representation

$$f(\Theta, x) = \sum_{q\in Q} e^{2iq\cdot x} f^{(q)}(\Theta), \tag{6.17}$$

where $f^{(q)}(\Theta)$ does not depend on x, whence

$$\partial_{x_i} f(\Theta, x) = \sum_{q\in Q} 2iq_i e^{2iq\cdot x} f^{(q)}(\Theta), \tag{6.18}$$

so that, for any $x \in \mathcal{X}$,

$$\begin{aligned} |\partial_{x_i} f(\Theta, x)| &\leq 2 \sum_{q\in Q} |q_i| \left| f^{(q)}(\Theta) \right| \\ &\leq 2 \left(\sum_{q\in Q} |q_i|^2 \right)^{1/2} \left(\sum_{q\in Q} \left| f^{(q)}(\Theta) \right|^2 \right)^{1/2} \\ &\leq 2(2N_{\max}+1)^{\dim \mathcal{X}/2} N_{\max} \left(\int_{\mathcal{X}} f(\Theta, x')^2 \frac{dx'}{(\pi)^{\dim \mathcal{X}}} \right)^{1/2}. \end{aligned} \tag{6.19}$$

We notice that the final bound (6.19) does not depend on x. Calling $C = C(N_{\max}, \dim \mathcal{X}) := 2(2N_{\max}+1)^{\dim \mathcal{X}} N_{\max}^2$ and $d\mu(x) := dx/(\pi)^{\dim \mathcal{X}}$, which is a probability measure on $\mathcal{X}$, we therefore have that

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}} |\partial_{x_i} f(\Theta, x)|^2\right] \leq C \int_{\mathcal{X}} \mathbb{E}\left[(f(\Theta, x))^2\right] d\mu(x). \tag{6.20}$$

Let us consider $\Theta = \Theta_0$. For any $m \geq 1$ the functions $\mathbb{E}\left[(f(\Theta_0, x))^2\right]$ are continuous for $x \in \mathcal{X}$ and they converge uniformly to $\mathcal{K}(x, x)$ by (3.33). Then, on the compact domain $\mathcal{X}$, we are allowed to exchange the limit and the integral as follows:

$$\lim_{m\to\infty} \mathbb{E}\left[\sup_{x\in\mathcal{X}} |\partial_{x_i} f(\Theta_0, x)|^2\right] \leq C \int_{\mathcal{X}} \lim_{m\to\infty} \mathbb{E}\left[(f(\Theta_0, x))^2\right] d\mu(x) = C \int_{\mathcal{X}} \mathcal{K}(x, x) d\mu(x)$$

Being $\mathcal{K}(x, x)$ the uniform limit of continuous functions, it is continuous too, so it has a maximum over its compact domain. This means that the above integral is finite. As a consequence, there exists a uniform bound M_0 on $\mathbb{E}\left[\sup_{x\in\mathcal{X}} |\partial_{x_i} f(\Theta_0, x)|^2\right]$:

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}} |\partial_{x_i} f(\Theta_0, x)|^2\right] \leq M_0 \qquad \forall m \geq 1. \tag{6.21}$$

Let us now generalize the previous result to $\Theta = \Theta_t$ for any $t \geq 0$. By Theorem 5.8, there exists $m_0 \in \mathbb{N}$ such that, with high probability,

$$\sup_{x\in\mathcal{X}} \left| f(\Theta_t, x) - f^{\text{lin}}(\Theta_t^{\text{lin}}, x) \right| \leq B(m) \qquad \text{with} \qquad \lim_{m\to\infty} B(m) = 0, \tag{6.22}$$

for all $m \geq m_0$ and for any $t \geq 0$, whence $B = \max_{m\geq m_0} B(m) < \infty$ and

$$|f(\Theta_t, x)| \leq |f^{\text{lin}}(\Theta_t^{\text{lin}}, x)| + B \qquad \forall\, x \in \mathcal{X}, \quad \forall\, m \geq m_0. \tag{6.23}$$

Furthermore, we know that for m large enough, $\hat{K}_{\Theta_0}(X, X^T)$ is invertible with high probability and therefore, by Lemma 5.17 and Remark 5.4 (see, in particular, (5.232)),

$$\begin{aligned} |f^{\text{lin}}(\Theta_t^{\text{lin}}, x) - f(\Theta_0, x)| &\leq \left| \hat{K}_{\Theta_0}(x, X^T) \hat{K}_{\Theta_0}^{-1} \left(\mathbb{1} - \left(\mathbb{1} - \eta_0 \hat{K}_{\Theta_0} t \right)^t \right) (F(0) - Y) \right| \\ &\leq \|\hat{K}_{\Theta_0}(x, X^T)\|_2 \|\hat{K}_{\Theta_0}^{-1}\|_F \|F(0) - Y\|_2. \end{aligned} \tag{6.24}$$

Both $\|\hat{K}_{\Theta_0}^{-1}\|_F$ and $\|F(0) - Y\|_2$ are bounded with high probability by Corollary 4.16 and Lemma 4.25 for m large enough; let their product be bounded by $\sqrt{C'}$ with high probability. We will denote with $\mathcal{E}$ the event in which all the previous bounds are satisfied. We can suppose that $\mathbb{P}(\mathcal{E}) \geq 1 - \delta/2$.
We will use a couple of times the following inequality. Let $X \geq 0$ be a non-negative random variable. Then

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{P}(\mathcal{E}) \mathbb{E}[\, X \mid \mathcal{E}\,] + \mathbb{P}(\mathcal{E}^c) \mathbb{E}\left[\, X \;\middle|\; \mathcal{E}^c \,\right] \\ &\geq \mathbb{P}(\mathcal{E}) \mathbb{E}[\, X \mid \mathcal{E}\,] \end{aligned} \tag{6.25}$$

Let us use (6.25) for this bound:

$$\mathbb{E}\left[(f(\Theta_0, x))^2 \;\middle|\; \mathcal{E} \right] \leq \frac{2}{2-\delta} \mathbb{E}\left[(f(\Theta_0, x))^2\right] \leq 2\mathcal{K}(x, x) \tag{6.26}$$

Let us now combine (6.20), (6.23) and (6.24):

$$\begin{aligned} &\limsup_{m\to\infty} \mathbb{E}\left[\sup_{\substack{x\in\mathcal{X} \\ t\geq 0}} |\partial_{x_i} f(\Theta_t, x)|^2 \;\middle|\; \mathcal{E} \right] \\ &\leq \limsup_{m\to\infty} 2C \int_{\mathcal{X}} \left(\mathbb{E}\left[\sup_{t\geq 0} \left(f^{\text{lin}}(\Theta_t, x) \right)^2 \;\middle|\; \mathcal{E} \right] + B^2 \right) d\mu(x) \\ &\leq \limsup_{m\to\infty} 2C \int_{\mathcal{X}} \Big(2\mathbb{E}\left[(f(\Theta_0, x))^2 \;\middle|\; \mathcal{E} \right] + B^2 \\ &\qquad\qquad + 2\mathbb{E}\left[\|\hat{K}_{\Theta_0}(x, X^T)\|_2^2 \|\hat{K}_{\Theta_0}^{-1}\|_F^2 \|F(0) - Y\|_2^2 \;\middle|\; \mathcal{E} \right] \Big) d\mu(x) \\ &\leq 2C \Big(4\mathcal{K}(x, x) + B^2 \\ &\qquad + 2C' \limsup_{m\to\infty} \int_{\mathcal{X}} \mathbb{E}\left[\|\hat{K}_{\Theta_0}(x, X^T)\|_2^2 \;\middle|\; \mathcal{E} \right] d\mu(x) \Big) \end{aligned} \tag{6.27}$$

Regarding the last term, we need a bound of the form

$$\limsup_{m\to\infty}\int_{\mathcal{X}}\mathbb{E}\left[\|\hat{K}_{\Theta_0}(x,X^T)\|_2^2\,\middle|\,\mathcal{E}\right]d\mu(x)\le A. \tag{6.28}$$

Let us recall that

$$\begin{aligned}\left(\hat{K}_{\Theta_0}(x,x')\right)^2&=\left(\nabla_\Theta f(\Theta_0,x)\cdot\nabla_\Theta f(\Theta_0,x')\right)^2\\&\le\|\nabla_\Theta f(\Theta_0,x)\|_2^2\|\nabla_\Theta f(\Theta_0,x')\|_2^2\\&\le\frac{1}{2}\|\nabla_\Theta f(\Theta_0,x)\|_2^2+\frac{1}{2}\|\nabla_\Theta f(\Theta_0,x')\|_2^2=\frac{1}{2}\hat{K}^2_{\Theta_0}(x,x)+\frac{1}{2}\hat{K}^2_{\Theta_0}(x',x')\end{aligned} \tag{6.29}$$

and write more explicitly

$$\begin{aligned}\mathbb{E}\left[\|\hat{K}_{\Theta_0}(x,X^T)\|_2^2\,\middle|\,\mathcal{E}\right]&=\sum_{j=1}^n\mathbb{E}\left[\hat{K}^2_{\Theta_0}(x,x^{(j)})\,\middle|\,\mathcal{E}\right]\\&=\frac{1}{2}\sum_{j=1}^n\left(\mathbb{E}\left[\hat{K}^2_{\Theta_0}(x,x)\,\middle|\,\mathcal{E}\right]+\mathbb{E}\left[\hat{K}^2_{\Theta_0}(x^{(j)},x^{(j)})\,\middle|\,\mathcal{E}\right]\right)\\&=\frac{n}{2}\mathbb{E}\left[\hat{K}^2_{\Theta_0}(x,x)\,\middle|\,\mathcal{E}\right]+\frac{1}{2}\sum_{j=1}^n\mathbb{E}\left[\hat{K}^2_{\Theta_0}(x^{(j)},x^{(j)})\,\middle|\,\mathcal{E}\right]\\&\le\frac{n}{2}\left(\mathrm{Var}[\,Y_0\,|\,\mathcal{E}\,]+(\mathbb{E}[\,Y_0\,|\,\mathcal{E}\,])^2\right)\\&\quad+\frac{1}{2}\sum_{j=1}^n\left(\mathrm{Var}[\,Y_j\,|\,\mathcal{E}\,]+\left(\mathbb{E}[\,Y_j\,|\,\mathcal{E}\,]\right)^2\right).\end{aligned} \tag{6.30}$$

where Y_j is the random variable

$$Y_j=\hat{K}_{\Theta_0}(x^{(j)},x^{(j)}),\qquad\text{with}\qquad x^{(0)}:=x. \tag{6.31}$$

By Assumption 4.3,

$$\lim_{m\to\infty}\sup_{x,x'\in\mathcal{X}}\left|\mathbb{E}[\hat{K}_{\Theta_0}(x,x')]-\bar{K}(x,x')\right|=0\qquad\forall x,x'\in\mathcal{X}, \tag{6.32}$$

i.e., $\mathbb{E}[Y_j]$ converges uniformly to $\bar{K}(x^{(j)},x^{(j)})$. Noticing that, by definition, $Y_j\ge 0$ for any $m>0$, we can use (6.25)

$$\mathbb{E}\left[Y_j\right]\ge\mathbb{P}(\mathcal{E})\mathbb{E}\left[\,Y_j\,|\,\mathcal{E}\,\right]\ge\left(1-\frac{\delta}{2}\right)\mathbb{E}\left[\,Y_j\,|\,\mathcal{E}\,\right],$$

whence

$$\mathbb{E}\left[\,Y_j\,|\,\mathcal{E}\,\right]\le\mathbb{E}\left[Y_j\right]+\frac{\delta}{2-\delta}\mathbb{E}\left[Y_j\right] \tag{6.33}$$

$$\limsup_{m\to\infty}\sup_{\substack{0\le j\le n\\x^{(0)}=x\in\mathcal{X}}}\left|\mathbb{E}\left[\,Y_j\,|\,\mathcal{E}\,\right]-\bar{K}(x^{(j)},x^{(j)})\right|\le\frac{\delta}{2-\delta}\sup_{x\in\mathcal{X}}\bar{K}(x,x)\le\sup_{x\in\mathcal{X}}\bar{K}(x,x) \tag{6.34}$$

Furthermore, being $\bar{K}(x,x')$ the uniform limit of continuous function, it is continuous, too; as a consequence, on its compact domain $\bar{K}(x,x')$ is bounded. Then, there is a constant G (which is uniform in $j=0,\dots,n$ and in $x^{(0)}=x\in\mathcal{X}$) such that

$$\mathbb{E}\left[\,Y_j\,|\,\mathcal{E}\,\right]\le G \tag{6.35}$$

for m large enough.

$$\limsup_{m\to\infty}\int_{\mathcal{X}}\left(\mathbb{E}[\,Y_j\,|\,\mathcal{E}\,]\right)^2d\mu(x)\le G^2 \tag{6.36}$$

We recall that in (4.215) we proved that (6.1) holds for the model $f(\Theta,x)$ as a function of the parameters, i.e., for any $x,x'\in\mathcal{X}$

$$\sup_{\substack{\Theta\in\mathscr{P}\\\theta'_k\in[0,2\pi)}}\left|\hat{K}_\Theta(x,x')-\hat{K}_{\Theta'}(x,x')\right|\le c_k\quad 1\le k\le Lm, \tag{6.37}$$

where $\Theta':=(\theta_1,\dots,\theta_{k-1},\theta'_k,\theta_{k+1},\dots,\theta_{Lm})$, with

$$c_k=16\,\frac{1}{N_K(m)}\,\frac{|\mathcal{M}||\mathcal{N}|}{N^2(m)}|\mathcal{M}_k|. \tag{6.38}$$

Furthermore, for any $m > 0$ and $x, x' \in \mathcal{X}$, the function $\Theta \mapsto \hat{K}_\Theta(x, x')$ is continuous on the compact domain $\mathscr{P} \ni \Theta$: it is therefore bounded and, as a consequence, square-integrable. Whence, by Lemma 6.1,

$$\mathrm{Var}[Y_j] \leq 64 \frac{Lm|\mathcal{M}^4|\mathcal{N}|^2}{N_K^2(m)N^4(m)} =: \Lambda(m). \tag{6.39}$$

Using again that $Y_j \geq 0$, we can bound

$$\begin{aligned}
\mathrm{Var}\left[\, Y_j\right] &= \mathbb{P}(\mathcal{E})\mathbb{E}\left[\, Y_j^2 \;\middle|\; \mathcal{E}\,\right] + \mathbb{P}(\mathcal{E}^c)\mathbb{E}\left[\, Y_j^2 \;\middle|\; \mathcal{E}^c\,\right] \\
&\quad + \left(\mathbb{P}(\mathcal{E})\mathbb{E}\left[\, Y_j \;\middle|\; \mathcal{E}\,\right] + \mathbb{P}(\mathcal{E}^c)\mathbb{E}\left[\, Y_j \;\middle|\; \mathcal{E}^c\,\right]\right)^2 \\
&\geq \left(\mathbb{P}(\mathcal{E})\right)^2 \left(\mathbb{E}\left[\, Y_j^2 \;\middle|\; \mathcal{E}\,\right] + \left(\mathbb{E}\left[\, Y_j \;\middle|\; \mathcal{E}\,\right]\right)^2\right) \\
&\geq \left(1 - \frac{\delta}{2}\right)^2 \mathrm{Var}\left[\, Y_j \;\middle|\; \mathcal{E}\,\right] \geq \frac{1}{4}\mathrm{Var}\left[\, Y_j \;\middle|\; \mathcal{E}\,\right]
\end{aligned} \tag{6.40}$$

Since $\lim_{m\to\infty} \Lambda(m) = 0$, we can compute

$$\limsup_{m\to\infty} \int_{\mathcal{X}} \mathrm{Var}\left[\, Y_j \;\middle|\; \mathcal{E}\,\right] d\mu(x) \leq 4 \limsup_{m\to\infty} \int_{\mathcal{X}} \mathrm{Var}\left[Y_j\right] d\mu(x) = \limsup_{m\to\infty} \Lambda(m) = 0. \tag{6.41}$$

So, for any $j = 0, \dots, n$,

$$\limsup_{m\to\infty} \int_{\mathcal{X}} \mathbb{E}\left[|\hat{K}_{\Theta_0}(x^{(j)}, x^{(j)})|^2\right] d\mu(x) \leq G^2, \tag{6.42}$$

whence

$$\limsup_{m\to\infty} \int_{\mathcal{X}} \mathbb{E}\left[\|\hat{K}_{\Theta_0}(x, X^T)\|_2^2\right] d\mu(x) \leq nG^2 =: A,$$

therefore there exists a constant $\tilde{M}$ which does not depend on i and t such that

$$\mathbb{E}\left[\sup_{\substack{x\in\mathcal{X}\\ t\geq 0}} |\partial_{x_i} f(\Theta_t, x)|^2 \;\middle|\; \mathcal{E}\right] \leq \tilde{M} \qquad \forall\, t \geq 0 \tag{6.43}$$

for any m large enough.

$$\limsup_{m\to\infty} \int_{\mathcal{X}} \mathbb{E}\left[|\hat{K}_{\Theta_0}(x, x)|^2\right] d\mu(x) \leq G^2 \tag{6.44}$$

Finally, by Markov's inequality

$$\mathbb{P}\left(|X| \geq a\right) \leq \frac{\mathbb{E}\left[|X|^2\right]}{a^2} \tag{6.45}$$

with

$$X = \sup_{\substack{x\in\mathcal{X}\\ t\geq 0}} |\partial_{x_i} f(\Theta_t, x)| \qquad \text{and} \qquad a = \sqrt{\frac{\tilde{M}}{2\delta}}, \tag{6.46}$$

we have

$$\mathbb{P}\left(\sup_{\substack{x\in\mathcal{X}\\ t\geq 0}} |\partial_{x_i} f(\Theta_t, x)| \geq \sqrt{\frac{\tilde{M}}{2\delta}} \;\middle|\; \mathcal{E}\right) \leq \frac{\delta}{2} \qquad \forall x \in \mathcal{X}, \quad 1 \leq i \leq \dim \mathcal{X}, \tag{6.47}$$

therefore the claim holds with probability at least $1 - \delta$. So far we proved that there exists $\bar{m} \in \mathbb{N}$ such that, with probability at least $1 - \delta$,

$$\max_{i=1,\dots,\dim\mathcal{X}} \sup_{x\in\mathcal{X}} |\partial_{x_i} f(\Theta_t, x)| \leq \bar{M} \tag{6.48}$$

for any $m \geq \bar{m}$. However, we know that

$$\sup_{\substack{\Theta\in\mathscr{P}, x\in\mathcal{X}\\ 1\leq i\leq \dim\mathcal{X}}} |\partial_{x_i} f(\Theta, x)| = M(m), \tag{6.49}$$

where, for any $m \in \mathbb{N}^*$, $M(m) < \infty$. Whence, if we set

$$M = \max\left\{M(1), \dots, M(\bar{m} - 1), \sqrt{\frac{\tilde{M}}{2\delta}}\right\}, \tag{6.50}$$

then

$$\sup_{\substack{x\in\mathcal{X}\\ t\geq 0}} |\partial_{x_i} f(\Theta_t, x)| \leq M \tag{6.51}$$

for any i, with probability at least $1 - \delta$, for all $m \geq 1$. □

Lemma 6.3 (Equiboundedness with high probability) *Let us suppose that the hypotheses of Lemma 6.2 are satisfied. Then, for any $\delta > 0$ and there exists $C > 0$ such that for any $m \geq 1$, with probability at least $1 - \delta$ we have*

$$\sup_{\substack{x \in \mathcal{X} \\ t \geq 0}} |f(\Theta_t, x)| \leq C. \tag{6.52}$$

Proof Let us fix $\bar{x} = x^{(1)} \in \mathcal{X}$. Let

$$\sigma_t^2 = \mathcal{K}_t(\bar{x}, \bar{x}), \quad \sigma^2 = \sup_{t \geq 0} \sigma_t^2, \quad \text{and} \quad \bar{y} = y^{(1)}. \tag{6.53}$$

By Lemma 5.18, it is easy to see that $\sigma^2 < \infty$. Furthermore, by Lemma 5.18, the random variable $f(\Theta_t, \bar{x})$ converges in distribution to $f_t^{(\infty)}(\bar{x})$, which are Gaussian random variables with mean $\mu_t(\bar{x})$ and variance σ_t^2. Given a constant R to be fixed later, we can compute

$$\begin{aligned} &\mathbb{P}\left[|f^{(\infty)_t}(\bar{x})) - \bar{y}| \geq R\right] \\ &\qquad = \mathbb{P}\left[f^{(\infty)}(\bar{x}) \geq R + \bar{y}\right] + \mathbb{P}\left[-f^{(\infty)}(\bar{x})) \geq R - \bar{y}\right], \end{aligned} \tag{6.54}$$

which can be estimated using Chernoff bound [61]:

$$\begin{aligned} \mathbb{P}\left[|f_t^{(\infty)}(\bar{x}) - \bar{y}| \geq R\right] &\leq \exp\left[-\frac{1}{2\sigma_t^2}(R + \bar{y})^2\right] + \exp\left[-\frac{1}{2\sigma_t^2}(R - \bar{y})^2\right] \\ &\leq 2\exp\left[-\frac{1}{2\sigma_t^2}(R - |\bar{y}|)^2\right]. \end{aligned} \tag{6.55}$$

Let

$$R(\delta, \bar{y}, \sigma^2) = \sqrt{2}\sigma\sqrt{\log\left(\frac{8}{\delta}\right)} + |\bar{y}|. \tag{6.56}$$

Then we can bound (6.55) as follows

$$\begin{aligned} \mathbb{P}\left[|f_t^{(\infty)}(\bar{x}) - \bar{y}| \geq R\right] &\leq 2\exp\left[-\frac{1}{2\sigma_i^2}\left(2\sigma^2 \log\left(\frac{8}{\delta}\right)\right)\right] \\ &\leq 2\exp\left[-\log\frac{8}{\delta}\right] = \frac{\delta}{4}, \end{aligned}$$

where we used that $\sigma^2/\sigma_t^2 \geq 1$. Since the random variable $f(\Theta_t, \bar{x})$ converges in distribution to $f_t^{(\infty)}(\bar{x})$, there exists $m_0 \in \mathbb{N}$ such that

$$\mathbb{P}\left[|f(\Theta_t, \bar{x}) - \bar{y}| \geq R\right] \leq 2\mathbb{P}\left[|f_t^{(\infty)}(\bar{x}) - \bar{y}| \geq R\right] \qquad \forall\, m \geq m_0 \tag{6.57}$$

because of Lemma 3.1. Therefore,

$$\mathbb{P}\left[|f(\Theta_t, \bar{x}) - \bar{y}| \geq R\right] \leq \frac{\delta}{2}. \tag{6.58}$$

Now, we have that, for any $x \in \mathcal{X}$

$$f(\Theta_t, x) = f(\Theta_t, \bar{x}) + \int_0^1 \left(\frac{d}{d\xi} f(\Theta_t, \xi x + (1 - \xi)\bar{x})\right) d\xi \tag{6.59}$$

whence, by Lemma 6.2, there is a constant $M(\delta/2)$

$$\begin{aligned} |f(\Theta_t, x)| &\leq |f(\Theta_t, \bar{x})| + \left|\int_0^1 (x - \bar{x}) \cdot \nabla_x f(\Theta_t, \xi x + (1 - \xi)\bar{x}) d\xi\right| \\ &\leq |f(\Theta_t, \bar{x})| + \|x - \bar{x}\|_\infty \left(\dim \mathcal{X} \sup_{1 \leq i \leq \dim \mathcal{X}} \sup_{x_0 \in \mathcal{X}} |\partial_{x_i} f(\Theta_t, x_0)|\right) \\ &\leq |f(\Theta_t, \bar{x})| + 2\pi \dim \mathcal{X}\, M\left(\frac{\delta}{2}\right) \end{aligned} \tag{6.60}$$

with probability at least $1 - \delta/2$. Therefore, for any $m \geq m_0$

$$|f(\Theta_t, x)| \leq R + |\bar{y}| + 2\pi \dim \mathcal{X} M = \tilde{C} \tag{6.61}$$

with probability at least $1 - \delta$, where $\tilde{C}$ does not depend on x or m. However, as discussed at the end of the proof of Lemma 6.2, this result can be generalized as follows: for any δ there is a constant C such that

$$|f(\Theta_t, x)| \leq C \qquad \forall x \in \mathcal{X}, \forall t \geq 0 \tag{6.62}$$

with probability at least $1 - \delta$ for any $m \geq 1$, where $\tilde{C}$ does not depend on x or m. □

Theorem 6.4 (Convergence to a Gaussian process) *Let us consider any feature space $\mathcal{X} \subseteq [0,\pi]^{\dim \mathcal{X}}$, which may also be an infinite set. Let us suppose that*

$$\lim_{m\to\infty} \frac{L^2 m^2 |\mathcal{M}|^6 |\mathcal{N}|^4}{N^5(m)} \log N(m) = 0. \tag{6.63}$$

Then, for any fixed $t \geq 0$, as $m \to \infty$,

$$\{f(\Theta_t, \cdot)\}_{x\in\mathcal{X}} \xrightarrow{d} \{f_t^{(\infty)}(\cdot)\}_{x\in\mathcal{X}}, \tag{6.64}$$

where $\{f_t^{(\infty)}(\cdot)\}_{x\in\mathcal{X}}$ is the Gaussian process characterized by mean and covariance

$$\mu_t(x) = \bar{K}(x, X^T) K^{-1} \left(\mathbb{1} - \left(\mathbb{1} - \eta_0 \bar{K}\right)^t\right) Y \tag{6.65}$$

$$\begin{aligned}
\mathcal{K}_t(x,x') = \mathcal{K}_0(x,x'), \\
&- \bar{K}(x, X^T) \bar{K}^{-1} \left(\mathbb{1} - \left(\mathbb{1} - \eta_0 \bar{K}\right)^t\right) \mathcal{K}_0(X, x') \\
&- \bar{K}(x', X^T) \bar{K}^{-1} \left(\mathbb{1} - \left(\mathbb{1} - \eta_0 \bar{K}\right)^t\right) \mathcal{K}_0(X, x) \\
&+ \bar{K}(x, X^T) \bar{K}^{-1} \left(\mathbb{1} - \left(\mathbb{1} - \eta_0 \bar{K}\right)^t\right) \times \\
&\qquad \times \mathcal{K}_0(X, X^T) \left(\mathbb{1} - \left(\mathbb{1} - \eta_0 \bar{K}\right)^t\right) \bar{K}^{-1} \bar{K}(X, x').
\end{aligned} \tag{6.66}$$

Proof The hypothesis (6.63) is the same of Theorem 5.9, so by Lemma 5.16 we immediately see that the hypotheses of Lemma 6.2 and Lemma 6.3 are satisfied. We recall that both the parameter space and the model function depends on the number of qubits:

$$\Theta^{(m)} \in \mathscr{P} = \mathscr{P}^{(m)}, \qquad f(\Theta, x) = f^{(m)}(\Theta^{(m)}, x). \tag{6.67}$$

Let $\mu_{m,t}$ the probability distribution of the function generated by the quantum neural network with m qubits at time t; it is the probability distribution on $C^0(\mathcal{X}, \mathbb{R})$, the set continuous function from $\mathcal{X}$ to $\mathbb{R}$ (endowed with the sup norm), induced by the random initialization of the parameters of the circuit and, in case, by the statistical noise during the training. Let us fix $\delta > 0$. Then, by Lemma 6.2 and Lemma 6.3, there exist $M_\delta > 0$ and $C_\delta > 0$ such that, for any $m \geq 1$ and $t \geq 0$

$$\mathbb{P}\left(\sup_{\substack{x\in\mathcal{X} \\ 1\leq i\leq \dim\mathcal{X}}} |\partial_{x_i} f^{(m)}(\Theta_t, x)| \leq M_\delta, \quad \sup_{x\in\mathcal{X}} |f^{(m)}(\Theta_t, x)| \leq C_\delta \right) > 1 - \delta. \tag{6.68}$$

The family

$$K_\delta := \{g \in C^0(\mathcal{X}, \mathbb{R}) : |g(x) - g(y)| \leq M_\delta \dim \mathcal{X} \|x - y\|_\infty,\ |g(x)| \leq C_\delta \,\forall\, x, y \in \mathcal{X}\} \tag{6.69}$$

is a set of equicontinuous and equibounded continuous functions, so by Ascoli-Arzelà's theorem it is compact [68]. In particular, for any m and t,

$$\mu_{m,t}(K_\delta) > 1 - \delta \tag{6.70}$$

This means that, fixing $t \geq 0$, the family of measures $\{\mu_{m,t}\}_{m\geq 1}$ is tight. Then, by Prokhorov's theorem [69], each subsequence $\mu_{m(k),t}$ of $\{\mu_{m,t}\}_{m\geq 1}$ has a convergence subsubsequence $\mu_{m(k(n)),t}$ to a measure $\mu_{\infty,t}$ on $C^0(\mathcal{X}, \mathbb{R})$ according to the weak topology. This implies the convergence in distribution of the entire sequence $\{f^{(m)}(\Theta_t, \cdot)\}_{m\geq 1}$ to a unique stochastic process $\{g_t^{(\infty)}(x)\}_{x\in\mathcal{X}}$:

$$f^{(m)}(\Theta_t, \cdot) \xrightarrow{d} g_t^{(\infty)}(\cdot). \tag{6.71}$$

Let us characterize this stochastic process. Fixed any finite family $\mathcal{F}$ of inputs, by Theorem 5.9 we know that the limit distribution of $\{f(\Theta_t, x)\}_{x\in\mathcal{F}}$ is a multivariate Gaussian with mean function $\mu_t(x)$ and covariance $\mathcal{K}_t(x, x')$. Therefore, arbitrariness of F, by the limit distribution is

$$f^{(m)}(\Theta_t, \cdot) \xrightarrow{d} f_t^{(\infty)}(\cdot) \tag{6.72}$$

i.e., the unique limit is $\{g_t^{(\infty)}(x)\}_{x\in\mathcal{X}} \equiv \{f_t^{(\infty)}(x)\}_{x\in\mathcal{X}}$; this proves the claim. □

7 Conclusions

We have proved that in the limit of infinite width, quantum neural networks can always be trained in polynomial time with respect to the number of qubits as long as they do not suffer from barren plateaus, and that the probability distribution of the function generated by the trained network converges in distribution to a Gaussian process, whose mean and covariance can be computed analytically.

More precisely, we have proved the following results. First, the probability distribution of the function generated by a quantum neural network with randomly initialized parameters converges in distribution to a Gaussian process when the cardinalities of the light cones are sufficiently small (Theorem 3.7). Second, the trained model is able to perfectly fit the training set exponentially fast in time, and the probability distribution of the function generated by the trained network converges in distribution to a Gaussian process whose mean and covariance can be computed analytically (Theorem 4.15). We have also provided a quantitative bound (4.55) that shows that the training happens in the lazy regime (Theorem 4.10 and Theorem 4.12). Furthermore, we have taken into account the statistical noise due to the finite number of measurements to estimate the gradients of the cost function. In particular, we have studied the evolution of the model under gradient descent with an unbiased estimator of the gradient of the cost function (5.12). We have proved that a sufficiently large number of measurements ensures all the convergence results of the noiseless setting with high probability (Theorem 5.4 and Theorem 5.9). Such number grows polynomially with the number of qubits (Proposition 5.5), so that any quantum advantage is not hindered by the training procedure. From the mathematical point of view we also have given a novel contribution in this field: going beyond the standard assumption that the feature space is a finite set, we have rigorously proved that the convergence to a Gaussian process, both at initialization and during the training, is valid also when we consider an infinite feature space (Theorem 6.4).

For a real quantum circuit, the number of qubits will unlikely be large enough to completely suppress the non-Gaussianities in the probability distribution of the function generated by the quantum neural network. As in the case of classical neural networks, this should not be seen as a problem: on the contrary, the deviations from the Gaussian limit make the neural tangent kernel trainable and can help to improve the generalization power of the neural network [70]. However, a rigorous understanding of such deviations is still an open problem in both the classical and the quantum case.

We stress that, despite Ref. [19] conjectured that all the architectures for quantum neural networks that do not suffer from barren plateaus can be simulated efficiently on a classical computer, our results are valid also in regimes that have hope of quantum advantage. Indeed, from the discussion in subsection 2.5, a quantum neural network with the qubits arranged on a d-dimensional lattice and the two-qubit gates acting between neighboring qubits does not allow for naive efficient classical simulations whenever $d \geq 2$ and the depth grows at least logarithmically with the number of qubits. On the other hand, assuming that the variance of the function generated by the quantum neural networks considered in our paper behaves as in the architecture of Ref. [16][19], for any d the hypotheses of our theorems are always satisfied for a logarithmic scaling of the depth $L \simeq \epsilon \log m$ with ϵ small enough but independent on m.

Our results open several questions:

- Is it possible to determine quantitative bounds for the distance between the probability distribution of the function generated by a given trained quantum neural network with a finite number of qubits and the Gaussian process associated to the covariance at initialization and the analytic neural tangent kernel of the network? Answering this question would avoid the need to consider a sequence of quantum neural networks with increasing number of qubits and would determine the width at which the results of this paper become valid.
- How does the convergence of the probability distribution of the generated function to a Gaussian process depends on the size of the training set, *i.e.*, how should the number of qubits scale with the number of examples for our results to be valid? For classical neural networks, a width polynomial in the number of examples is enough [71], and we expect the same to hold in the quantum setting.
- Is it possible to prove nontrivial bounds to the normalization constant $N(m)$ of a generic quantum circuit and therefore determine whether the circuit suffers from barren plateaus? The study of the

[19] Formally, Ref. [16] considers random nonparametric two-qubit gates sampled from a 2-design. However, there are no reasons to believe that the architectures that we consider, where the two-qubit gates are fixed, will have a different scaling of the variance of the generated function.

detailed architecture of the class of circuits satisfying the hypotheses of our theorems would be essential to determine which architectures can have hopes of quantum advantage.

- Can the class of circuits satisfying of our theorems be extended? More precisely, we wonder whether there are circuits whose light cones grow faster than the requirements of the form (2.77) that appears in the hypotheses of our theorems, but whose output is trainable and converges to a Gaussian process. We do expect that the hypotheses of our theorems can be relaxed, but, in order to envisage a larger class of variational circuits for which our results hold, it may be necessary to formulate new proofs using different strategies.

All these questions will be considered in our future works, with the hope to answer to some of them and to give new insights on the working principles of quantum neural networks. Finally, we hope that our results can open the way to answering the main open problem in quantum machine learning: are there problems of practical relevance that variationally trained quantum neural networks can solve better than any classical algorithm?

A A counterexample for large light cones

In this section, we will show that, if we relax the hypothesis of Theorem 3.7 concerning the growth of the light cones, we can exhibit a circuit which generates a function whose distribution at initialization is not a Gaussian process.

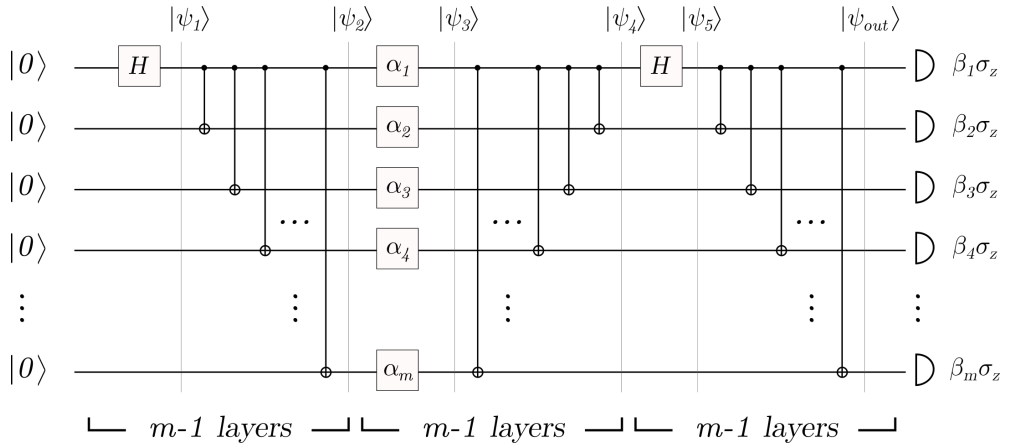


Fig. 23 Pathological circuit

Let us consider the circuit in the figure. It is a variational circuit in which the feature encoding gates and most parameter gates are trivial, since they act as the identity on the qubits. For this reason they are not represented in the figure. The only exception is the layer $\ell = m$, in which we identify $\alpha_k = \theta_{[mk]}$. Each layer has one C-NOT as an entangling gate and, in the layers $\ell = 1$ and $\ell = 2m - 1$, a Hadamard gate appears. The total number of layers is $L = 3m - 3 = O(m)$ and the light cones have cardinality $O(m)$. The local observables are Pauli Z rescaled by a bounded factor $0 \leq \beta_k \leq 1$.

The initial state, after the Hadamard gate, becomes

$$|00\dots0\rangle \xmapsto{H_1} |\psi_1\rangle \frac{1}{\sqrt{2}} \left(|00\dots0\rangle + |10\dots0\rangle\right). \tag{A.1}$$

After the first $m-1$ layers the state becomes

$$|\psi_2\rangle = \frac{1}{\sqrt{2}} \left(|00\dots0\rangle + |11\dots1\rangle\right) \tag{A.2}$$

The single qubit gates labelled by α_k are phase gates

$$P(\alpha_i) = \begin{pmatrix} 1 & \\ & e^{i\alpha_i} \end{pmatrix} \qquad \text{in the computational basis.} \tag{A.3}$$

The state after the action of these gates is

$$|\psi_3\rangle = \frac{1}{\sqrt{2}} \left(|00\ldots0\rangle + e^{i\sum_{k=1}^{m}\alpha_k} |11\ldots1\rangle \right). \tag{A.4}$$

If we assume that α_k are i.i.d. random variables distributed as

$$\mathbb{P}(\alpha_k = 0) = \frac{1}{2}, \qquad \mathbb{P}(\alpha_k = \pi) = \frac{1}{2}, \tag{A.5}$$

we have, by symmetry,

$$\mathbb{P}\left(\sum_{k=1}^{m} \alpha_k \equiv 0 \bmod 2\pi \right) = \mathbb{P}\left(\sum_{k=1}^{m} \alpha_k \equiv \pi \bmod 2\pi \right) = \frac{1}{2}. \tag{A.6}$$

So the following states are equally likely

$$\frac{1}{\sqrt{2}} \left(|00\ldots0\rangle \pm |11\ldots1\rangle \right). \tag{A.7}$$

The following C-NOTs produce the state

$$|\psi_4\rangle = \frac{1}{\sqrt{2}} \left(|00\ldots0\rangle \pm |10\ldots0\rangle \right) = \frac{1}{\sqrt{2}} (|0\rangle \pm |1\rangle) \otimes |0\ldots0\rangle. \tag{A.8}$$

Then, by the Hadamard gate,

$$|\psi_4\rangle = \frac{1}{\sqrt{2}} (|0\rangle \pm |1\rangle) \otimes |0\ldots0\rangle \xmapsto{H_1} |\psi_5\rangle = \begin{cases} |0\rangle \otimes |0\ldots0\rangle \\ |1\rangle \otimes |0\ldots0\rangle \end{cases}. \tag{A.9}$$

The final C-NOTs yield

$$|\psi_5\rangle = \begin{cases} |0\rangle \otimes |0\ldots0\rangle \\ |1\rangle \otimes |0\ldots0\rangle \end{cases} \quad \mapsto \quad |\psi_{out}\rangle = \begin{cases} |00\ldots0\rangle \\ |11\ldots1\rangle \end{cases}. \tag{A.10}$$

The local observables are given by

$$O_k = \beta_k \sigma_z^{(k)} \qquad \text{with} \qquad \beta_k = \sqrt{k} - \sqrt{k-1}. \tag{A.11}$$

Then, the output function is

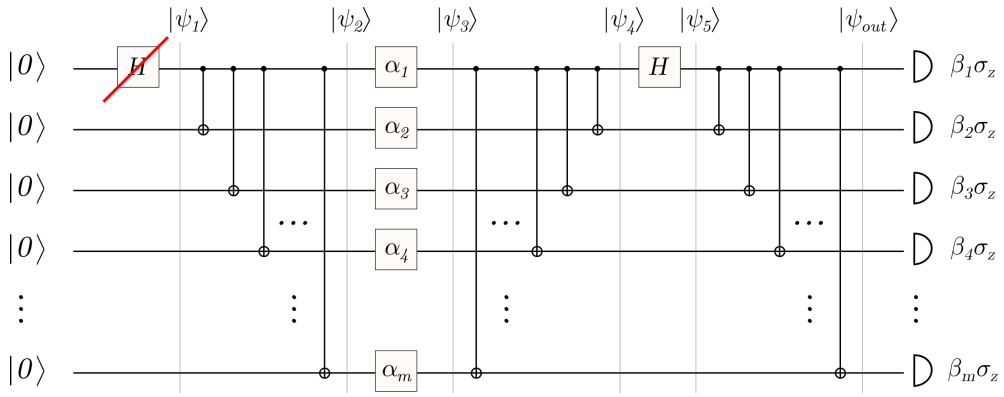
$$\begin{aligned} f(\Theta) &= \frac{1}{N(m)} \langle \psi_{out} | \mathcal{O} | \psi_{out} \rangle \qquad \text{with} \qquad N(m) = \sqrt{m} \\ &= \frac{\pm 1}{\sqrt{m}} \big(\sqrt{1} + (\sqrt{2} - \sqrt{1}) + (\sqrt{3} - \sqrt{2}) + \cdots + (\sqrt{m} - \sqrt{m-1}) \big) \\ &= \pm 1, \end{aligned} \tag{A.12}$$

which means that

$$\mathbb{P}(f(\Theta) = 1) = \mathbb{P}(f(\Theta) = -1) = \frac{1}{2}. \tag{A.13}$$

This distribution does not converge to a Gaussian process as $m \to \infty$.

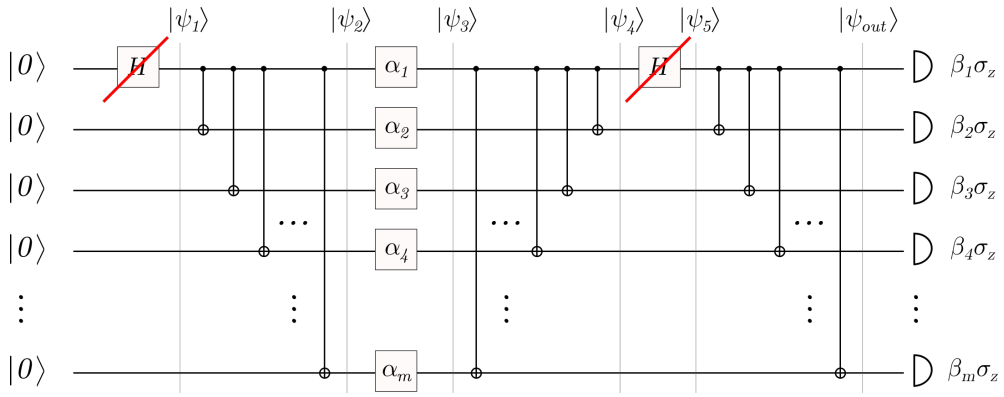
This circuit has a remarkable property: if we perturb the first qubit by a local operation which cancels the first Hadamard gate, the result is completely changed and the result does not depend on the parameters α_i.



$$|\psi_1\rangle = |\psi_2\rangle = |\psi_3\rangle = |\psi_4\rangle = |00\cdots0\rangle\,, \qquad |\psi_5\rangle = |+\rangle \otimes |0\cdots0\rangle\,, \tag{A.14}$$

$$|\psi_{out}\rangle = \frac{1}{\sqrt{2}}(|00\cdots0\rangle + |11\cdots1\rangle). \tag{A.15}$$

If we also cancel the second Hadamard gate, the output changes again: it is equal to the input.



$$|\psi_1\rangle = |\psi_2\rangle = |\psi_3\rangle = |\psi_4\rangle = |\psi_5\rangle = |\psi_{out}\rangle = |00\cdots0\rangle\,. \tag{A.16}$$

In both cases, a "small" perturbation of the circuit (i.e., a unitary transformation acting on a single qubit) has important consequences in the final state of all the qubits. A more robust behaviour of the output state with respect to such perturbations would be a very desirable property for a variational circuit.

B Architecture-independent bounds for light cones

Lemma B.1 (Architecture-independent bounds are exponential in L) *Without any knowledge on the family of interactions $\mathcal{I}_U$ of a circuit U, we can estimate the cardinalities of the light cones as follows:*

$$|\mathcal{N}_k| \leq 2^{L+1}, \tag{B.1}$$

$$|\mathcal{M}_{[\ell k]}| \leq 2^{L+1-\ell}. \tag{B.2}$$

Proof Since we use only two-qubit gates, $|\mathcal{I}_{\ell,k}| \leq 2$.
By (2.21) and (2.22), we have

$$|\mathcal{N}_k^\ell| = |\mathcal{J}_k^\ell| \leq \begin{cases} |\mathcal{I}_{L,k}| \leq 2 & \ell = L, \\ |\mathcal{J}_k^{\ell+1}| \max_{k'} |\mathcal{I}_{\ell,k'}| \leq 2|\mathcal{J}_k^{\ell+1}| & \ell < L. \end{cases} \tag{B.3}$$

So, inductively,

$$\begin{cases} |\mathcal{N}_k^\ell| \leq 2|\mathcal{N}_k^{(\ell+1)}| \\ |\mathcal{N}_k^L| \leq 2 \end{cases} \quad \rightarrow \quad \left|\mathcal{N}_k^{(L-\ell)}\right| \leq 2^{\ell+1}. \tag{B.4}$$

Whence, by Definition 2.5,

$$|\mathcal{N}_k| \leq \sum_{\ell=0}^{L-1} |\mathcal{N}_k^{(L-\ell)}| \leq \sum_{\ell=1}^{L} 2^\ell \leq 2^{L+1}, \tag{B.5}$$

which is (B.1). Now, by Lemma 2.3, if $\ell < L$,

$$|\mathcal{M}_{[\ell k]}| \leq |\mathcal{I}_{\ell,k}| \max_{k'} |\mathcal{M}_{[(\ell+1)\, k']}| \leq 2 \max_{k'} |\mathcal{M}_{[(\ell+1)\, k']}|. \tag{B.6}$$

and, since

$$\mathcal{M}_{[Lk]} = \mathcal{I}_{L,k} \quad \rightarrow \quad |\mathcal{M}_{[Lk]}| \leq 2, \tag{B.7}$$

we have the recurrence relation

$$\begin{cases} \max_k |\mathcal{M}_{[\ell k]}| \leq 2 \max_k |\mathcal{M}_{[(\ell+1)\, k]}| & \ell < L \\ \max_k |\mathcal{M}_{[Lk]}| \leq 2 \end{cases}. \tag{B.8}$$

Therefore

$$|\mathcal{M}_{[\ell k]}| \leq \max_{\bar{k}} |\mathcal{M}_{[\ell \bar{k}]}| \leq 2^{L-\ell+1}. \tag{B.9}$$

□

Corollary B.2 *The bounds of Lemma B.1 imply the following bounds on the maximal cardinalities and the sums defined above*

$$|\mathcal{M}| \leq 2^L \qquad |\mathcal{N}| \leq 2^{L+1} \qquad \Sigma_n \leq 2m2^{nL} \tag{B.10}$$

Proof The first two bounds are immediate, while the third can be proved as follows:

$$\begin{aligned} \Sigma_n &= \sum_{i=1}^{Lm} |\mathcal{M}_i|^n = \sum_{k=1}^{m} \sum_{\ell=1}^{L} |\mathcal{M}_{[\ell k]}|^n \leq \sum_{k=1}^{m} \sum_{\ell=1}^{L} \left(2^{L+1-\ell}\right)^n \\ &= m \sum_{\ell=1}^{L} \left(2^\ell\right)^n \leq m \sum_{\ell=0}^{L} \left(2^n\right)^\ell = m \frac{(2^n)^{L+1} - 1}{2^n - 1} \\ &\leq \left(\frac{2^n}{2^n - 1}\right) m 2^{nL} \leq 2m2^{nL}. \end{aligned} \tag{B.11}$$

□

C Computation of the analytic NTK

In this appendix we explain how $K(x, x') := \mathbb{E}\left[\hat{K}_\Theta(x, x')\right]$ can be analytically computed for any given circuit with finite width. Using the fundamental rules of quantum mechanics, a laborious but elementary linear algebra computation allows to write $f(\Theta, x)$, and therefore $\hat{K}_\Theta(x, x')$, as an analytic function of Θ for any fixed x and x'. Clearly, this naive method does not provide a classically efficient algorithm, since it is based on computations in the local Hilbert spaces of each observable; such spaces typically have superpolynomial dimension with respect to the number of qubits. In the spirit of the proof of Lemma 4.30, we already know which kind of dependence on Θ will appear: since $\hat{K}_\Theta(x, x')$ is defined in terms of the derivatives of $f(\Theta, x)$ and $f(\Theta, x')$, by virtue of (4.178) we will obtain an expression of the form

$$\hat{K}_\Theta(x, x') = \frac{1}{N_K(m)} \left(\sum_{v \in \{0, \pm 1\}^{Lm}} \tilde{f}_v(x) e^{2i\Theta \cdot v} 2iv \right) \cdot \left(\sum_{w \in \{0, \pm 1\}^{Lm}} \tilde{f}_w(x') e^{2i\Theta \cdot w} 2iw \right) \tag{C.1}$$

$$= \frac{1}{N_K(m)} \sum_{v,w\in\{0,\pm1\}^{Lm}} \tilde{f}_v(x)\tilde{f}_w(x')e^{2i\Theta\cdot(v+w)}(-4v\cdot w), \tag{C.2}$$

$$\tag{C.3}$$

where the coefficients $f_v(\,\cdot\,)$ are the results of the linear algebra computations in the local Hilbert spaces of each observable. Therefore, we see that the computation of the expectation value at initialization is elementary as well, since the parameters are uniform random variables (see Assumption 4.4):

$$K(x,x') = \mathbb{E}\left[\hat{K}_\Theta(x,x')\right] \tag{C.4}$$

$$= \frac{-4}{N_K(m)} \sum_{v,w\in\{0,\pm1\}^{Lm}} \tilde{f}_v(x)\tilde{f}_w(x')(v\cdot w)\prod_{j=1}^{Lm}\mathbb{E}_{\theta_j}\left[e^{2i(v_j+w_j)\theta_j}\right] \tag{C.5}$$

$$= \frac{-4}{N_K(m)} \sum_{v,w\in\{0,\pm1\}^{Lm}} \tilde{f}_v(x)\tilde{f}_w(x')(v\cdot w)\prod_{j=1}^{Lm}\mathbb{1}_{v_j=-w_j} \tag{C.6}$$

$$= \frac{4}{N_K(m)} \sum_{v\in\{0,\pm1\}^{Lm}} \tilde{f}_v(x)\tilde{f}_{-v}(x')(v\cdot v), \tag{C.7}$$

where $\mathbb{1}_{v_j=-w_j}$ is the characteristic function

$$\mathbb{1}_{v_j=-w_j} = \begin{cases} 1 & v_j = -w_j \\ 0 & v_j \neq -w_j \end{cases}. \tag{C.8}$$

This shows that an analytical form for $K(x,x')$ can be computed.

We stress that the algorithm outlined above will in general have a runtime that grows polynomially in the dimension of the local Hilbert spaces of Definition 2.53 and will not be efficient. However, our results are not meant as an alternative algorithm to replace the sequence of quantum neural networks, but just as a means to mathematically characterize the probability distribution of the functions generated by the trained quantum neural networks. Indeed, if the covariance at initialization and the analytic neural tangent kernel could be computed efficiently on a classical computer, the associated kernel regression algorithm would also be efficient to run on a classical computer and therefore quantum neural networks would have no advantage in the regime where our results apply.

Declarations

Funding and/or Conflicts of interests/Competing interests

GDP has been supported by the HPC Italian National Centre for HPC, Big Data and Quantum Computing - Proposal code CN00000013 - CUP J33C22001170001 and by the Italian Extended Partnership PE01 - FAIR Future Artificial Intelligence Research - Proposal code PE00000013 - CUP J33C22002830006 under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU. Funded by the European Union - NextGenerationEU under the National Recovery and Resilience Plan (PNRR) - Mission 4 Education and research - Component 2 From research to business - Investment 1.1 Notice Prin 2022 - DD N. 104 del 2/2/2022, from title "understanding the LEarning process of QUantum Neural networks (LeQun)", proposal code 2022WHZ5XH – CUP J53D23003890006. GDP is a member of the "Gruppo Nazionale per la Fisica Matematica (GNFM)" of the "Istituto Nazionale di Alta Matematica "Francesco Severi" (INdAM)".
No funding was received by FG for conducting this study.

The authors have no competing interests to declare that are relevant to the content of this article.

Data availability

We do not analyse or generate any datasets, because our work proceeds within a theoretical and mathematical approach.

Acknowledgements

We thank Dario Trevisan for useful discussions.

References

[1] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., *et al.*: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015) https://doi.org/10.1038/nature14236

[2] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015) https://doi.org/10.1038/nature14539

[3] Radford, A., Metz, L., Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks (2016)

[4] Schmidhuber, J.: Deep learning in neural networks: An overview. Neural networks **61**, 85–117 (2015)

[5] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press. http://www.deeplearningbook.org (2016)

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is All you Need. Curran Associates, Inc. (2017). https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[7] Cerezo, M., Arrasmith, A., Babbush, R., Benjamin, S.C., Endo, S., Fujii, K., McClean, J.R., Mitarai, K., Yuan, X., Cincio, L., Coles, P.J.: Variational quantum algorithms. Nature Reviews Physics **3**(9), 625–644 (2021) https://doi.org/10.1038/s42254-021-00348-9

[8] Schuld, M., Sweke, R., Meyer, J.J.: Effect of data encoding on the expressive power of variational quantum-machine-learning models. Physical Review A **103**(3) (2021) https://doi.org/10.1103/physreva.103.032430

[9] Havlíček, V., Córcoles, A.D., Temme, K., Harrow, A.W., Kandala, A., Chow, J.M., Gambetta, J.M.: Supervised learning with quantum-enhanced feature spaces. Nature **567**(7747), 209–212 (2019) https://doi.org/10.1038/s41586-019-0980-2

[10] Schuld, M., Killoran, N.: Quantum machine learning in feature hilbert spaces. Physical Review Letters **122**(4) (2019) https://doi.org/10.1103/physrevlett.122.040504

[11] Liu, Y., Arunachalam, S., Temme, K.: A rigorous and robust quantum speed-up in supervised machine learning. Nature Physics **17**(9), 1013–1017 (2021) https://doi.org/10.1038/s41567-021-01287-z

[12] You, X., Wu, X.: Exponentially Many Local Minima in Quantum Neural Networks. PMLR (2021). https://proceedings.mlr.press/v139/you21c.html

[13] Anschuetz, E.R., Kiani, B.T.: Quantum variational algorithms are swamped with traps. Nature Communications **13**(1) (2022) https://doi.org/10.1038/s41467-022-35364-5

[14] McClean, J.R., Boixo, S., Smelyanskiy, V.N., Babbush, R., Neven, H.: Barren plateaus in quantum neural network training landscapes. Nature Communications **9**(1), 4812 (2018) https://doi.org/10.1038/s41467-018-07090-4

[15] Cerezo, M., Sone, A., Volkoff, T., Cincio, L., Coles, P.J.: Cost function dependent barren plateaus in shallow parametrized quantum circuits. Nature Communications **12**(1) (2021) https://doi.org/10.1038/s41467-021-21728-w

[16] Napp, J.: Quantifying the barren plateau phenomenon for a model of unstructured variational ansätze (2022)

[17] Kiani, B.T., Palma, G.D., Marvian, M., Liu, Z.-W., Lloyd, S.: Learning quantum data with the quantum earth mover's distance. Quantum Science and Technology **7**(4), 045002 (2022) https://doi.org/10.1088/2058-9565/ac79c9

[18] Palma, G.D., Marvian, M., Trevisan, D., Lloyd, S.: The quantum wasserstein distance of order 1. IEEE Transactions on Information Theory **67**(10), 6627–6643 (2021) https://doi.org/10.1109/tit.2021.3076442

[19] Cerezo, M., Larocca, M., García-Martín, D., Diaz, N.L., Braccia, P., Fontana, E., Rudolph, M.S., Bermejo, P., Ijaz, A., Thanasilp, S., Anschuetz, E.R., Holmes, Z.: Does provable absence of barren plateaus imply classical simulability? Or, why we need to rethink variational quantum computing (2023)

[20] Neal, R.M.: Priors for Infinite Networks, pp. 29–53. Springer, New York, NY (1996). https://doi.org/10.1007/978-1-4612-0745-0_2 . https://doi.org/10.1007/978-1-4612-0745-0_2

[21] Lee, J., Bahri, Y., Novak, R., Schoenholz, S.S., Pennington, J., Sohl-Dickstein, J.: Deep Neural Networks as Gaussian Processes (2018)

[22] Lee, J., Xiao, L., Schoenholz, S.S., Bahri, Y., Novak, R., Sohl-Dickstein, J., Pennington, J.: Wide neural networks of any depth evolve as linear models under gradient descent*. Journal of Statistical Mechanics: Theory and Experiment **2020**(12), 124002 (2020) https://doi.org/10.1088/1742-5468/abc62b

[23] Hanin, B.: Which Neural Net Architectures Give Rise To Exploding and Vanishing Gradients? (2018)

[24] Hanin, B.: Neural Networks and Large Width and Depth (lecture notes of the mini-course given by the author in the Rome Center on Mathematics for Modeling and Data Science at Tor Vergata in January 2023) (2023)

[25] Hanin, B.: Correlation Functions in Random Fully Connected Neural Networks at Finite Width (2022)

[26] Hanin, B.: Random Neural Networks in the Infinite Width Limit as Gaussian Processes (2021)

[27] Yang, G.: Scaling Limits of Wide Neural Networks with Weight Sharing: Gaussian Process Behavior, Gradient Independence, and Neural Tangent Kernel Derivation (2020)

[28] Yang, G.: Tensor Programs I: Wide Feedforward or Recurrent Neural Networks of Any Architecture are Gaussian Processes (2021)

[29] Yang, G.: Tensor Programs II: Neural Tangent Kernel for Any Architecture (2020)

[30] Yang, G., Littwin, E.: Tensor Programs IIb: Architectural Universality of Neural Tangent Kernel Training Dynamics (2021)

[31] Yang, G.: Tensor Programs III: Neural Matrix Laws (2021)

[32] Yang, G., Hu, E.J.: Feature Learning in Infinite-Width Neural Networks (2022)

[33] Yang, G., Hu, E.J., Babuschkin, I., Sidor, S., Liu, X., Farhi, D., Ryder, N., Pachocki, J., Chen, W., Gao, J.: Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer (2022)

[34] Yang, G., Littwin, E.: Tensor Programs IVb: Adaptive Optimization in the Infinite-Width Limit (2023)

[35] Yang, G., Yu, D., Zhu, C., Hayou, S.: Tensor Programs VI: Feature Learning in Infinite-Depth Neural Networks (2023)

[36] Larocca, M., Ju, N., García-Martín, D., Coles, P.J., Cerezo, M.: Theory of overparametrization in quantum neural networks. Nature Computational Science **3**(6), 542–551 (2023) https://doi.org/10.1038/s43588-023-00467-6

[37] Liu, J., Tacchino, F., Glick, J.R., Jiang, L., Mezzacapo, A.: Representation learning via quantum neural tangent kernels. PRX Quantum **3**, 030323 (2022) https://doi.org/10.1103/PRXQuantum.3.030323

[38] Abedi, E., Beigi, S., Taghavi, L.: Quantum lazy training. Quantum **7**, 989 (2023) https://doi.org/10.22331/q-2023-04-27-989

[39] Liu, J., Najafi, K., Sharma, K., Tacchino, F., Jiang, L., Mezzacapo, A.: Analytic theory for the dynamics of wide quantum neural networks. Phys. Rev. Lett. **130**, 150601 (2023) https://doi.org/10.1103/PhysRevLett.130.150601

[40] García-Martín, D., Larocca, M., Cerezo, M.: Deep quantum neural networks form Gaussian processes (2023)

[41] You, X., Chakrabarti, S., Chen, B., Wu, X.: Analyzing Convergence in Quantum Neural Networks: Deviations from Neural Tangent Kernels (2023)

[42] Seeger, M.: Pac-bayesian generalisation error bounds for gaussian process classification. Journal of machine learning research **3**(Oct), 233–269 (2002)

[43] Rad, A.: Deep Quantum Neural Networks are Gaussian Process (2023)

[44] Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning series. MIT Press, London, England (2005)

[45] Vaart, A.W.v.d.: Asymptotic Statistics. Cambridge University Press (1998). https://doi.org/10.1017/CBO9780511802256

[46] Kallenberg, O.: Gaussian and Poisson Convergence, pp. 125–146. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-61871-1_7 . https://doi.org/10.1007/978-3-030-61871-1_7

[47] Williams, D.: Probability with Martingales. Cambridge University Press (1991). https://doi.org/10.1017/CBO9780511813658

[48] Fristedt, B.E., Gray, L.F.: A Modern Approach to Probability Theory, 1997 edn. Probability and Its Applications. Birkhauser Boston, Secaucus, NJ (1996)

[49] Döring, H., Jansen, S., Schubert, K.: The Method of Cumulants for the Normal Approximation (2021)

[50] Féray, V., Méliot, P.-L., Nikeghbali, A.: Dependency graphs and mod-Gaussian convergence, pp. 95–110. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46822-8_9 . https://doi.org/10.1007/978-3-319-46822-8_9

[51] Hogg, R., Craig, A.T.: Introduction to Mathematical Statistics, 6th edn. Pearson, Upper Saddle River, NJ (2004)

[52] Romano, J.P., Siegel, A.F.: Counterexamples in probability and statistics. Routledge (2017)

[53] Janson, S.: Normal Convergence by Higher Semiinvariants with Applications to Sums of Dependent Random Variables and Random Graphs. The Annals of Probability **16**(1), 305–312 (1988) https://doi.org/10.1214/aop/1176991903

[54] Döring, H., Eichelsbacher, P.: Moderate deviations via cumulants. Journal of Theoretical Probability **26**(2), 360–385 (2012) https://doi.org/10.1007/s10959-012-0437-0

[55] Robbins, H.: A remark on stirling's formula. The American Mathematical Monthly **62**(1), 26–29 (1955). Accessed 2023-07-30

[56] Brandao, F.G., Harrow, A.W.: Product-state approximations to quantum states. Communications in Mathematical Physics **342**, 47–80 (2016)

[57] Karimi, H., Nutini, J., Schmidt, M.: Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In: Frasconi, P., Landwehr, N., Manco, G., Vreeken, J. (eds.) Machine Learning and Knowledge Discovery in Databases, pp. 795–811. Springer, Cham (2016)

[58] Marrero, C.O., Kieferová, M., Wiebe, N.: Entanglement Induced Barren Plateaus (2021)

[59] You, X., Chakrabarti, S., Wu, X.: A Convergence Theory for Over-parameterized Variational Quantum Eigensolvers (2022)

[60] Jacot, A., Gabriel, F., Hongler, C.: Neural Tangent Kernel: Convergence and Generalization in Neural Networks (2020)

[61] Chernoff, H.: A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations. The Annals of Mathematical Statistics **23**(4), 493–507 (1952) https://doi.org/10.1214/aoms/1177729330

[62] Schuld, M., Bergholm, V., Gogolin, C., Izaac, J., Killoran, N.: Evaluating analytic gradients on quantum hardware. Physical Review A **99**(3) (2019) https://doi.org/10.1103/physreva.99.032331

[63] Giaquinta, M., Modica, G.: Mathematical Analysis: An Introduction to Functions of Several Variables. Birkhäuser Boston (2010). https://books.google.it/books?id=0YE_AAAAQBAJ

[64] McDiarmid, C.: On the method of bounded differences. Cambridge University Press (1989). https://doi.org/10.1017/CBO9781107359949.008

[65] Slutsky, E.: Über stochastische Asymptoten und Grenzwerte. Metron 5, Nr. 3, 3-89 (1925). (1925)

[66] Chizat, L., Oyallon, E., Bach, F.: On Lazy Training in Differentiable Programming (2020)

[67] Boucheron, S., Lugosi, G., Massart, P.: Concentration Inequalities. Oxford University Press, London, England (2013)

[68] Rudin, W.: Principles of Mathematical Analysis, 3rd edn. International series in pure and applied mathematics. McGraw-Hill Professional, New York, NY (1976)

[69] Billingsley, P.: Convergence of Probability Measures, 2nd edn. Wiley Series in Probability and Statistics. John Wiley & Sons, Nashville, TN (1999)

[70] Misiakiewicz, T., Montanari, A.: Six Lectures on Linearized Neural Networks (2023)

[71] Arora, S., Du, S.S., Hu, W., Li, Z., Salakhutdinov, R.R., Wang, R.: On Exact Computation with an

Infinitely Wide Neural Net. Curran Associates, Inc. (2019). https://proceedings.neurips.cc/paper_files/paper/2019/file/dbc4d84bfcfe2284ba11beffb853a8c4-Paper.pdf